PERGAMON

Vision Research 38 (1998) 2401–2428

Vision Research

Training 'greeble' experts: a framework for studying expert object recognition processes

Isabel Gauthier [a,*], Pepper Williams [b], Michael J. Tarr [c], James Tanaka [d]

[a] *Department of Psychology, Yale University, P.O. Box 208205, New Haven, CT 06520, USA*
[b] *University of Massachusetts, Boston, MA, USA*
[c] *Brown University, Providence, RI, USA*
[d] *Oberlin College, Oberlin, OH, USA*

Received 10 January 1997; received in revised form 29 May 1997; accepted 21 November 1997

Abstract

Twelve participants were trained to be experts at identifying a set of 'Greebles', novel objects that, like faces, all share a common spatial configuration. Tests comparing expert with novice performance revealed: (1) a surprising mix of generalizability and specificity in expert object recognition processes; and (2) that expertise is a multi-faceted phenomenon, neither adequately described by a single term nor adequately assessed by a single task. Greeble recognition by a simple neural-network model is also evaluated, and the model is found to account surprisingly well for both generalization and individuation using a single set of processes and representations. © 1998 Elsevier Science Ltd. All rights reserved.

Keywords: Configural encoding; Face recognition; Neural networks; Object categorization; Perceptual expertise

1. Introduction

Are the mechanisms used by perceivers as they become increasingly familiar with an object class the same as those used by perceivers when they first encounter the class? To date, expertise in object recognition has been discussed primarily in relation to face recognition, and much of this discussion has focused on the hypothesis that faces are recognized through a specialized brain module, separate from the system used to recognize other classes of objects [1]. However, two factors distinguish face recognition from recognition of other classes of objects [2]: first, we recognize faces at a subordinate, rather than the basic, level [3], and second, we are experts at recognizing faces, while we are not expert recognizers of most other types of object. Although expert recognition processes may be most often applied to faces, and novice, basic-level categorization may be sufficient for many of our interactions with other types of objects, there are many contexts in which observers must discriminate among individual exemplars of non-face object classes. For instance, salespeople need to distinguish between different models of cars, shoes, or sports rackets; lumberjacks need to be able to recognize individual trees, and fighter pilots need to identify different types of planes. These individuals will all have extensive experience with such categories (i.e. they are experts), but other people with less experience will still be able to make the same kind of discriminations, even if they do so less efficiently than experts.

Therefore, individual face recognition may be seen as but the most salient example of an expert, subordinate-level recognition problem [2]. Other such problems, which vary along a continuum of difficulty and for which observers vary along a continuum of expertise, may be approachable in a qualitatively similar way. Two basic hypotheses underly our study: One is that expertise may mediate many apparent dissociations that occur in visual recognition (such as that between face and object recognition) and the second is that tuning of a single kind of visual representation can subserve both novice and expert object recognition performance, regardless of the object class.

In this paper, we build on a training method first developed by Gauthier and Tarr [4] to explore in more detail the process of expertise acquisition. Twelve par-

* Corresponding author. E-mail: isabel.gauthier@yale.edu.

PII: S0042-6989(97)00442-2

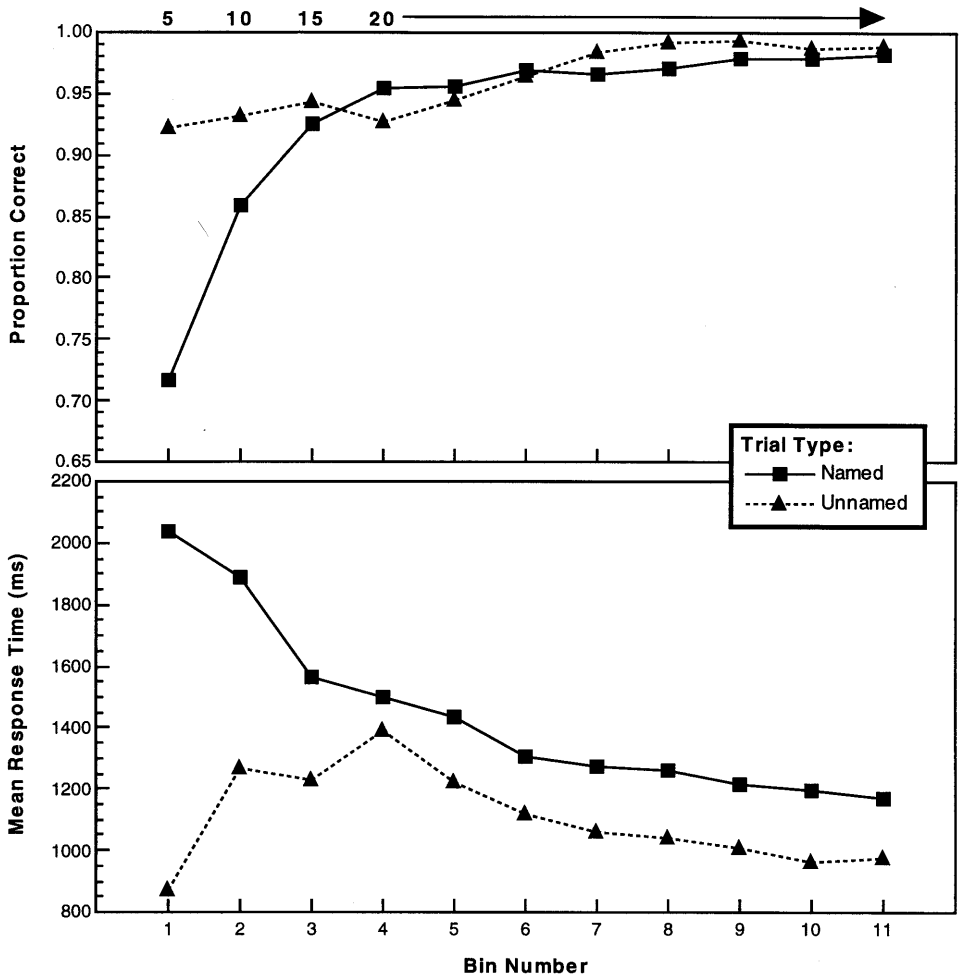

Fig. 1. Greebles can be grouped into two different 'genders' (PLOKs and GLIPs) and several different 'families'.

ticipants were trained for approximately nine hours each to become expert recognizers of a novel class of object called Greebles (Fig. 1). The performance of these experts on several tasks known or suspected to be affected by expertise was compared to novices' performance on the tasks. Section 3 examines learning names for new (not seen during training) sets of Greebles, Sections 4 and 5 use tests designed to reveal 'configural' and/or 'holistic' [5] processing, and Section 6 tests for possible limits on experts' Greeble recognition abilities. Finally, we present a simple neural-network model that demonstrates how a single recognition system can subserve both individual- and class-level recognition and lead to some of the expertise effects found in the present study using the same processes and representations.

All Greebles share similar parts in a common spatial configuration, so like faces, Greebles are only distinguishable on the basis of 'second-order properties' [6], i.e. configurations of parts. Furthermore, the testing conditions we adopted were motivated in large part by findings from the face recognition literature. Therefore, findings that putative face-specific effects can be obtained with Greebles may be taken as evidence that the visual system neither requires nor possesses a module specifically dedicated to face recognition *per se*. Beyond the question of how faces are recognized, we hope to address the more general issue of how the visual object recognition system adapts to more efficiently process any overlearned class of homogeneous stimuli that must be individuated.

2. Expert training

Because so few experimental studies of perceptual expertise have been reported, little is known about the best methods for manipulating the level of expertise. It is obvious that experts are generally more experienced than novices, but it is not clear exactly how much experience is necessary to produce significant 'expertise effects'. It is also unclear how one should deal with individual differences in the rate of learning.

There are essentially two types of training methods that one can adopt: all participants can receive a fixed amount of training that the experimenters hope will be enough to result in expertise, or each participant can receive a variable amount of training, until he or she reaches a pre-specified criterion. Gauthier and Tarr [4] used the latter strategy, training participants until they were as fast to recognize Greebles at the individual level as they were to recognize them at a more categorical level. This criterion was derived from a study by Tanaka and Taylor [7], who found that expert dog judges and birdwatchers were as fast at recognizing different types of dogs and birds at the subordinate level as they were at recognizing dogs and birds at the categorical level (whereas novices were much faster to say that an object was a 'bird' than that it was a 'sparrow'). An important advantage of this procedure is that if we accept the assumption that the criterion is an adequate measure of expertise level, we can be reasonably confident that at the conclusion of training, all participants will be experts. Also, experts' performance may be more meaningfully compared across studies that manipulate various aspects of the training procedure, again assuming that participants reach the criterion only once they are experts, and assuming that all aspects of expertise are equally assessed by the criterion.

Unfortunately, there is always a danger that some participants will reach the criterion, and thus their training will be terminated, before they are actually experts. Furthermore, the criterion training method makes it somewhat difficult to assess the results of tasks performed during the training procedure itself, since different participants will go through different amounts of training. Therefore, in the present study, we elected to have each participant perform exactly the same training procedure. However, we used Gauthier and Tarr's results [4] as a guide, training all of our participant;s longer than the average time necessary for participants in Gauthier and Tarr's study to reach the expertise criterion. We were also able to assess if and when each participant in our study would have met the expertise criterion adopted by Gauthier and Tarr.

Section 2 presents the method and results of the present training procedure. Besides training for a set amount of time rather than to a criterion, our procedure differed in several other important ways from the one used by Gauthier and Tarr [4]. First, whereas Gauthier and Tarr trained participants to classify Greebles at three different levels (gender, family, and individual; see Fig. 1), we only included gender- and individual-level training. Second, we required participants to discriminate 20 different Greebles at the individual level, as opposed to ten in Gauthier and Tarr's study. Finally, the present procedure used a more heterogeneous group of training tasks: Gauthier and Tarr had participants perform the same Verification task (see below for a description of this task) over and over for the bulk of the training, while we required participants to alternate between the Verification task and a Naming task during most of the training sessions. The latter two changes were intended to make the training more difficult, under the assumption that the harder participants had to work during training, the more expertise they would develop with Greebles.

2.1. Method

2.1.1. Participants

Twelve undergraduates from Oberlin College participated in the experiment in return for cash payment.

2.1.2. Materials

The stimuli were 30 Greebles [4], photorealistically-rendered 3D objects that all share a common configuration. Each Greeble is made up of a vertically-oriented 'body' with four protruding 'appendages', from top to bottom, two 'boges', a 'quiff' and a 'dunth'. As shown in Fig. 1, Greebles can be easily categorized into two different classes, which we will refer to as 'genders', on the basis of the orientation of appendages (up or down), and into five different 'families' on the basis of the shape of the main body. These two categorical dimensions are orthogonal, rather than hierarchical: Each individual Greeble is a member of one gender and one family, and is distinguishable from other members of its gender and family by the shapes of its appendages. Every appendage is unique in the set, although some pairs are more similar than others. The two genders and 20 of the individual Greebles (ten of each gender and four of each family) were given nonsense-word labels (e.g. 'vali' or 'pimo'); each label started with a different letter. The Greebles were all rendered with the same purple shade, stippled texture, and overhead lighting direction. Images were about 6.5 cm high $\times$ 3.25 cm wide, and when viewed from about 60 cm from the screen, yielded a display area of approximately $6.2 \times 3.1°$ of visual angle. The experiments were conducted on Macintosh computers equipped with color monitors (72 pixels per in.).

2.1.3. Procedure

The training procedure required participants to learn and then practice recognizing the Greebles at both the gender and individual levels. In all, each participant was trained for approximately 9 h, during ten 1-h sessions spread out over 2 weeks. Each session included some combination of seven different tasks:

1. *Gender inspect*—participants saw Greebles along with their gender labels, and made no response.
2. *Gender categorization*—participants saw Greebles without labels, and pressed the 'P' key for 'ploks' or the 'G' key for 'glips'.

3. *Individual inspect*—participants saw Greebles with their names and made no response.
4. *Naming with response*—participants saw Greebles with their names and pressed the key corresponding to the first letter of the name.
5. *Naming with feedback*—participants saw Greebles without names, attempted to press the correct key to name the Greeble, and if incorrect, saw the Greeble again, this time with its correct name.
6. *Naming*—participants saw Greebles alone and attempted to name them.
7. *Verification*—participants saw either a gender label or an individual name for 1000 ms, then after a pause of 200 ms saw a Greeble, and responded with one key if they thought the Greeble matched the label/name, or another key if they thought the Greeble and label/name did not match.

Every trial in all tasks was preceded by a fixation cross shown for 250 ms, and in tasks that required a response, participants always heard a 'beep' when they responded incorrectly.

The specific order of tests in each session is shown in Table 1. Note that participants learned the gender labels and names for five individual Greebles in the first session, then learned five more Greebles in each of the following three sessions. In the individual inspect, naming with response, and naming with feedback tests, participants only saw Greebles for which they had learned names. In the naming and verification tasks, however, all 30 training Greebles were shown, even if participants did not know names for them. The correct response for unnamed Greebles in the naming task was to press the space bar (designated the NIL response). Each Greeble was shown twice in this task, for a total of 60 trials. When unnamed Greebles were seen in the verification task, participants were to respond 'same' if they were preceded by a NIL label, or 'different' if they were preceded by the name of another Greeble. Thus there were seven different types of verification trials: a gender label could be followed by a Greeble from that gender or a different one; an individual name could be followed by the appropriate Greeble, a Greeble that was known by another name, or an unnamed Greeble; or a NIL label could be followed by an unnamed or a named Greeble. Each Greeble was seen four times in the verification task, once each following the correct gender label, the incorrect gender label, the correct individual name (or NIL), and the incorrect individual name (or NIL).

2.1.4. Analyses

Data from the verification and naming tasks were analyzed for trends over the course of training. Since results varied fairly widely from test to test for individual participants, we combined groups of tests into 'bins'. Bin 1 included naming tests # 1–2 and verifica-

Table 1
Training procedure

Trials	Task
Session 1	
1	Gender inspect (ten Greebles at once)
10	Gender inspect
30	Gender categorization
10	Individual inspect (set 1[a])
10	Naming with response (set 1)
15	Naming with feedback (set 1)
60	Naming
120	Verification
30	Gender categorization
10	Naming with response (set 1)
60	Naming
120	Verification
476	Total trials this session
Session 2	
6	Gender inspect
20	Naming with response (set 1)
120	Verification
10	Individual inspect (set 2)
10	Naming with response (set 2)
30	Naming with feedback (sets 1/2)
60	Naming
120	Verification
736	Total trials this session
Session 3	
6	Gender inspect
40	Naming with response (sets 1/2)
120	Verification
10	Individual inspect (set 3)
10	Naming with response (set 3)
45	Naming with feedback (sets 1/2/3)
60	Naming
120	Verification
771	Total trials this session
Session 4	
6	Gender inspect
60	Naming with response (sets 1/2/3)
120	Verification
10	Individual inspect (set 4)
10	Naming with response (set 4)
60	Naming with feedback (sets 1/2/3/4)
60	Naming
120	Verification
806	Total trials this session
Sessions 5, 6, 7, 8, 9	
40	Naming with response (sets 1/2/3/4)
60	Naming
120	Verification
760	Total trials these sessions
Session 10	
40	Naming with response (sets 1/2/3/4)
60	Naming
120	Verification

Table 1 (continued)

Trials	Task
220	Total trials this session

[a] Sets 1, 2, 3, and 4 refer to the first, second, third, and fourth group of 5 Greebles whose individual names were learned during training. See text for descriptions of test tasks.

tion tests #1–3, bin 2 included naming tests #3–5 and verification tests #4–7, bin 3 included naming tests #6–8 and verification tests #8–11, and subsequent bins included three naming and three verification tests. These bins were constructed so that: (a) there were an equivalent number (11) of bins for the naming and verification tests; (b) there were approximately the same number of tests (2–3 for naming and 3–4 for verification) in each bin, and (c) bins 1, 2, 3 and 4 included tests for which participants had to identify 5, 10, 15 and 20 different Greebles, respectively, at the individual level. That is, participants learned the first five individual Greebles names just before the first test in bin 1, learned the second five names just before the first test in bin 2, and learned the third and fourth set of names just before the first tests in bins 3 and 4, respectively. Accuracy and response times, averaged across tests in each bin, were analyzed for each individual participant and for all participants combined. Unless otherwise noted, all response times reported in this paper are geometric means (which are less susceptible to the effects of outliers than are arithmetic means), calculated on correct trials only. An alpha level of 0.05 was adopted for all inferential statistics, and significance levels are only reported for marginally significant (between 0.05 and 0.10) tests.

2.2. Results and discussion

2.2.1. Naming task

Group means for naming test performance are shown in Fig. 2. 'Named' trials (solid lines and squares in the graph) are those on which participants saw a Greeble for which they knew a name, and were required to press the key corresponding to the first letter of the name. For 'unnamed' trials (dashed lines and triangles in the graph), participants did not know a name for the Greeble, and pressed the space bar (for the NIL response) to indicate this fact. Participants were strikingly good at naming Greebles. By the fourth bin (which included tests from the fourth training session), at which point participants were performing a 20-key naming task, participants were naming Greebles in about 1500 ms per trial, and achieving greater than 95% accuracy. Furthermore, ten of our 12 experts returned 8–13 weeks after the last training session for further testing, and received two Naming with Feedback tests for the 20 named training Greebles. Remarkably, experts achieved a 58% accuracy rate on the first test, and 85% on the second test; chance performance would have been 5%.

Note that the proportion of named and unnamed trials varied during the first four bins of training. In the first bin of tests, five Greebles were named and 25 unnamed; just before the first test in the second, third, and fourth bins, participants learned names for five previously unnamed Greebles. Thus, each of these four bins had progressively more Greebles that required participants to retrieve names and progressively fewer Greebles that required a NIL response. Intuitively, it might seem that these changes would result in more difficult naming judgments, since more Greebles must be individually distinguished from one another, and less difficult NIL responses, since fewer Greebles require this response. Contrary to this intuition, participants became progressively faster and more accurate at making name responses and slower at making NIL responses in this period. These trends were confirmed by a significant interaction between trial type and bin for these four bins, $F(3, 33) = 31.2$ for response time and $F(3, 33) = 19.9$ for accuracy. Performance on both types of trials increased (i.e. response times got faster) from the fourth bin on. These effects were also apparent in most of the individual participants' results, as seen in Fig. 3.

This pattern of results probably reflects a combination of several factors. As more and more Greebles were distinguished as individuals, it became more likely that an unnamed Greeble would be similar to a named one, and thus more difficult to label as NIL. Of course, it also became more likely that a named Greeble would be similar to another named Greeble. Apparently, this

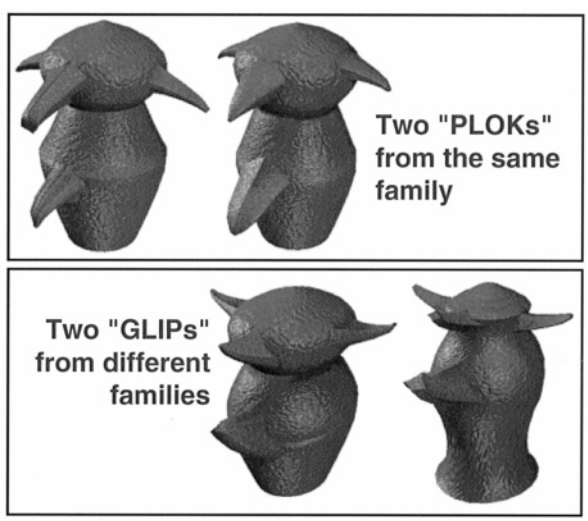

Fig. 2. Performance on the naming task throughout training. Numbers above the accuracy graph indicate the number of Greebles participants knew the names for in each bin of tests.

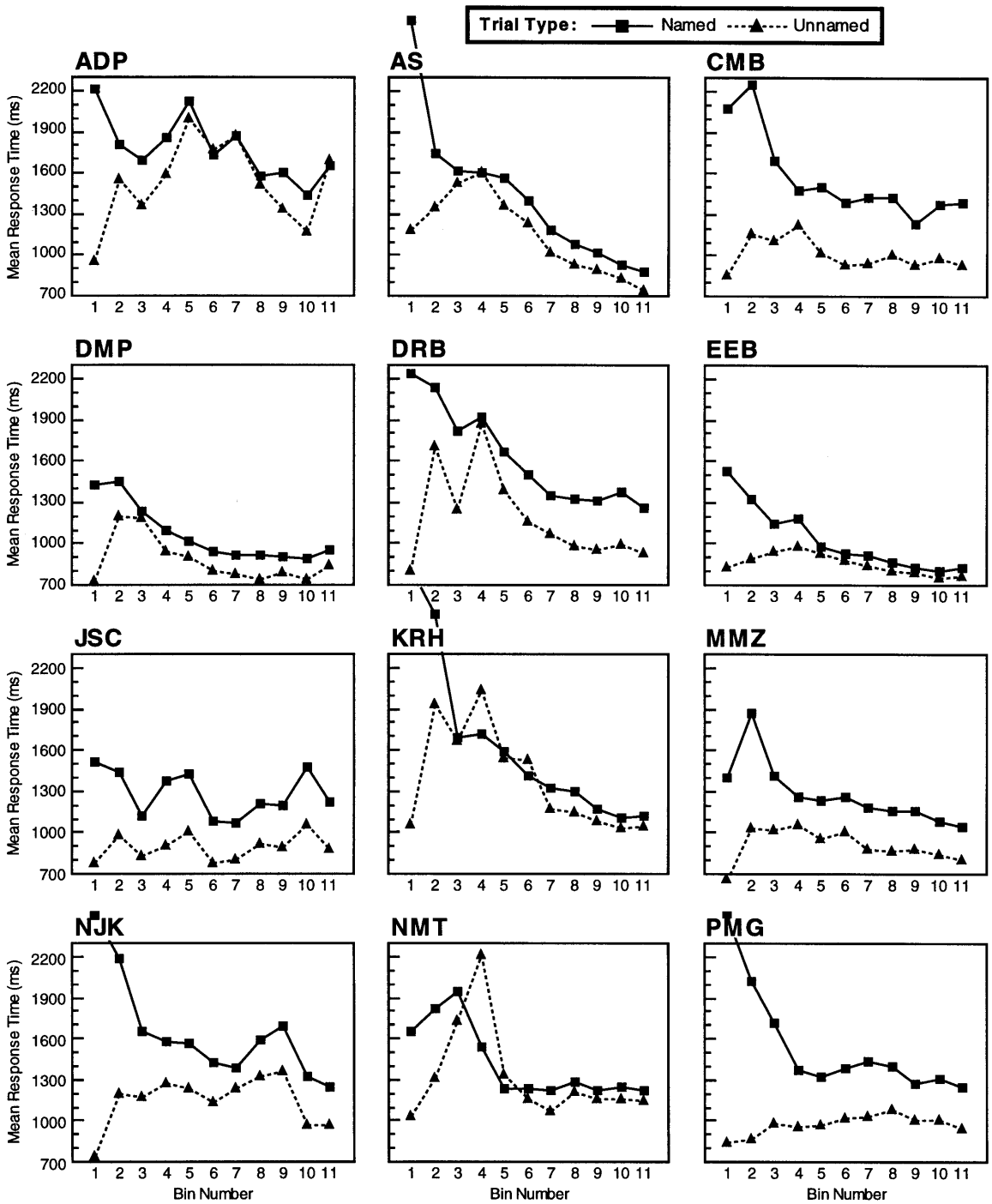

Fig. 3. Performance on the naming task throughout training for each expert participant.

cost was outweighed by the benefits gained from increased fluency with Greebles whose names were already practiced. Participants may have also been learning individuating information about unnamed Greebles even though they were associating all of these Greebles with the same response. This would hurt performance on Greebles that still required NIL responses, since any familiar Greeble would seem to require a name. However, it could help participants with newly-named Greebles, because these Greebles would already have been somewhat familiar to participants at the time when their names were being acquired (this would also account for why performance on unnamed Greebles got better from the fourth bin on, when no more new names had to be learned).

If this were the case, then Greebles whose names were learned later in the training procedure should have been easier to learn to name. A post-hoc analysis of the

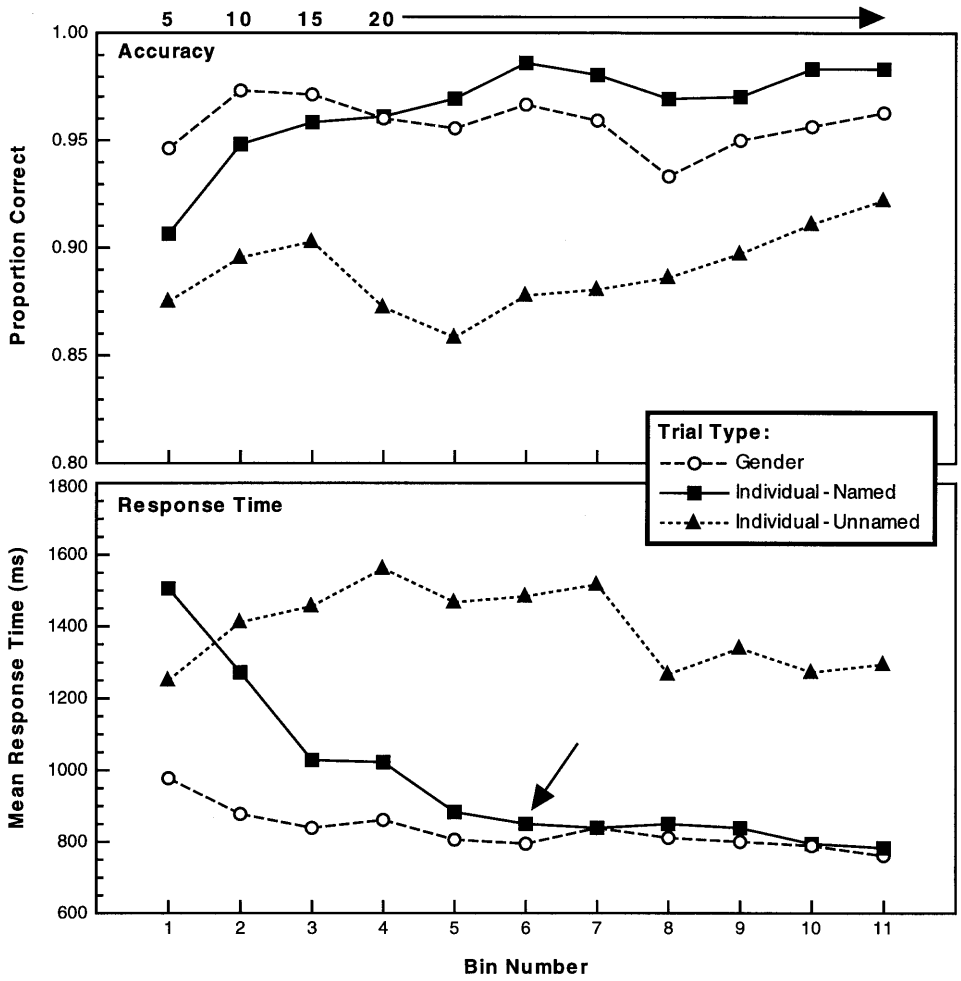

Fig. 4. Performance on 'yes' trials in the verification task throughout training. Numbers above the accuracy graph indicate the number of Greebles participants knew names for in each bin of tests. As indicated by the arrow on the RT graph, the significance level of the difference between Gender and individual-named trials, as computed by post-hoc LSD tests, was not significant for the 6th bin onward.

naming data supports this prediction: Mean accuracy for the first set of named Greebles in the first bin of tests was 0.72; while accuracy for the second, third, and fourth sets of named Greebles, which were initially named in the second, third, and fourth bins of tests, was 0.78, 0.91 and 0.91, respectively. These differences were significant, $F(3, 33) = 16.2$. However, since we did not plan this analysis, we did not properly counterbalance the order in which participants learned to name the Greebles, so at least part of the effect could be due to idiosyncrasies in which Greebles were included in each set. Furthermore, improvement in name-learning with training could also have been due to more general 'learning to learn'—i.e. participants could have learned what distinguishing information to look for when forced to individuate new Greebles. Nevertheless, this analysis suggests that learning effects in the absence of identification may be an interesting avenue for future research in the expertise-training paradigm.

2.2.2. Verification tests

Group means for 'same' trials on the verification test are shown in Fig. 4, and individual means for each participant in Fig. 5. These are trials in which participants saw a gender label (circles and long-dashed lines in the figure), an individual name (squares and solid lines), or the NIL label (triangles and short-dashed lines), then saw a Greeble that matched this label. Results for corresponding 'different' trials are consistent with these data, but are left off the graphs for clarity.

Looking first at the relationship between individual-named and individual-unnamed trials, we see some important similarities and differences with the naming task. As was the case with naming, over the first four bins of trials, participants got progressively faster at matching names with correctly-named Greebles, but progressively slower at matching the NIL label with unnamed Greebles. This pattern was significant, $F(3, 33) = 21.59$, and probably reflects the same mechanisms discussed in connection with the Naming task. When

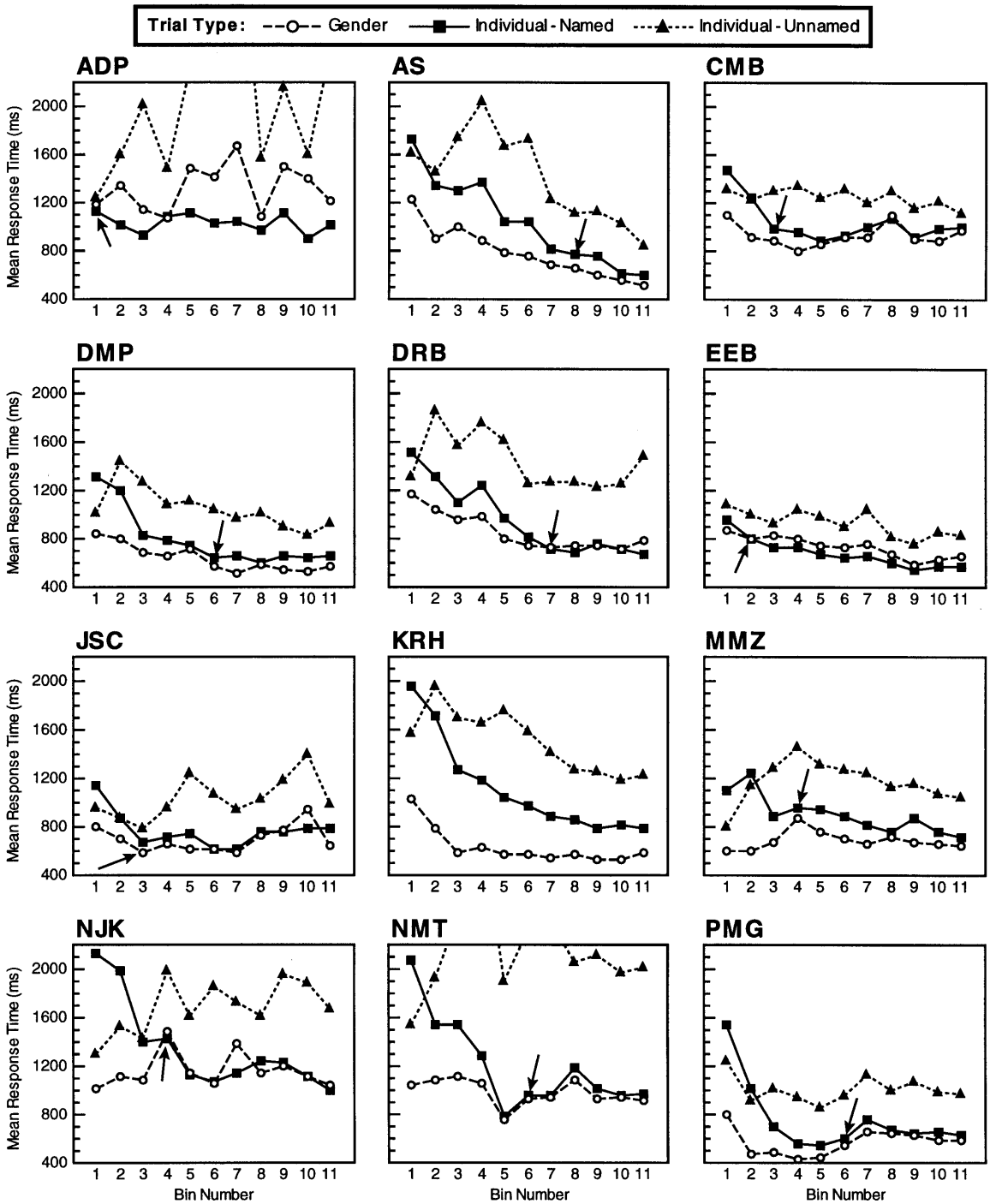

Fig. 5. Performance on 'yes' trials in the verification task throughout training for each expert participant. The arrow on each graph indicates the first bin on which the participant's performance on individual-named trials was not significantly different from performance on gender trials.

naming Greebles, however, participants were almost always faster to make NIL decisions than name decisions, whereas when verifying the match between labels and Greebles, participants were almost always slower for NIL than for named trials. This dissociation reflects an important difference between the two tasks. When given a name in the verification task, participants can generate an expectation for the Greeble they may subsequently see. If the Greeble matches this expectation, they can respond 'same'; otherwise, they respond 'different'. However, it is difficult to generate an expectation from the NIL label, since so many different Greebles are assigned to this label.

During the first bin of tests, participants were presumably able to quickly reject unnamed Greebles as any of the five that they knew names for, and as a

result, ten of the 12 participants were faster on NIL trials than on named trials. As more Greebles became labeled over the following three bins, however, this process became more time-consuming, and participants quickly became slower on NIL than on named trials. Interestingly, seven participants (AS, CMB, DMP, DRB, EEB, KRH, and MMZ) showed a pattern of gradual improvement on NIL trials from at least the fourth bin onward, while performance for the other five participants on the NIL trials declined throughout the course of training. Four of these five participants also failed to improve on NIL trials in the naming task (the exception was NMT, who improved dramatically between bins 4 and 5, but then stayed at roughly this same level throughout the remainder of the training). This correspondence leads us to suspect that participants used one of two different strategies for dealing with unnamed Greebles. Perhaps the non-improvers continued with the strategy of going through each of the named Greebles before concluding that a Greeble was unnamed, whereas the improvers learned to explicitly associate all of the unnamed Greebles with the NIL label and response. The former strategy could explain why some participants were actually worse in later bins than earlier ones. As unnamed Greebles become more familiar, it would become more difficult to determine that it is not a named Greeble.

Turning to the relationship between individual-named and gender trials, we see that nearly all of the participants were at some point as fast to verify a Greeble's individual name as they were to verify a Greeble's gender (or at least not reliably slower), and thus would have reached the criterion used by Gauthier and Tarr [4] for completion of training. To assess exactly when participants in our experiments would have passed this criterion, we performed ANOVAs including the factors of bin number (1–11) and trial type (individual-named and gender) on both the group data and on each individual participant's data. Arrows in Figs. 4 and 5 indicate the bin on which the significance level of LSD tests computed from these ANOVAs first rose above the 0.05 level.[1]

A more stringent criterion could also require that performance on the Gender trials be at an asymptotic level. Applying such a restriction to the present data would shift the criterion by at least one bin for participants EEB, MMZ, and NJK. Participant KRH never met our criterion, and showed signs of asymptoting in the last three bins, meaning that even given further training, he might not have reached the criterion. Participant ADP produced a pattern of response times remarkably different than any of the other 11 experts: He was significantly faster on individual-named than on gender trials in the very first bin, but showed no systematic improvement in response times for any of the trial conditions. Moreover, although his overall hit rate (averaged across all bins) for named Greebles was not abnormally low (0.94, as compared to an average of 0.97 across all participants), his false alarm rate (responding 'same' when the correct answer was 'different') for these Greebles was 0.19, whereas the average for all participants was 0.06. His performance on Gender trials, both when the correct response was 'yes' or 'no,' was also atrocious. Interestingly, however, his performance pattern in subsequent tests (described in Sections 3–6) did not differ qualitatively from that of other experts. His data was therefore included in all subsequent analyses.

2.2.3. Verification–naming correlations

Data from the verification tests indicate a quantitative effect of expertise training: Over the course of the ten training sessions, the difference in mean response time between verifying a Greeble's gender and verifying a Greeble's individual identity systematically decreased from an average of just over 500 ms to essentially 0. While this quantitative shift alone indicates a certain amount of expertise with the Greebles, one could argue that perceptual expertise should be defined not only by this overall increase in performance but also by a qualitative shift in the type of information used to recognize an object. Did our participants show such a qualitative shift in the course of becoming experts? One indicator would be a change in the relative difficulty with which individual Greebles were recognized. That is, if the information used by participants to identify Greebles changed over the course of the training procedure, then different Greebles might be expected to be easy or hard to identify at the beginning than at the end of training. To get more stable means for individual Greebles, we performed this analysis on four larger groups of six to eight tests each. We also limited the analysis to the first two sets of Greebles (ten individuals) for which names were learned, since the final set was not learned until the fourth session, at which point many participants had already reached (or were very close to reaching) the expertise criterion. All reported correlations are on response times, since accuracy levels were very high throughout training.

When means for each Greeble were computed across all 12 participants, the Pearson product–moment correlation between name verification performance on the first and last group of tests was a very large 0.75, whereas the correlation between naming performance on these two groups of tests was only 0.41. Looking at each individual participant's data separately, eight of the 12 participants showed a larger correlation for the

[1] Note that we were actually being conservative in this analysis by using the LSD test, since a more stringent post-hoc test would have resulted in quicker estimates of when participants reached the expert level.

verification than for the Naming test. Interestingly, the correlation between performance on the two tests was initially very high (0.91 in the first group of tests), but decreased dramatically over the course of training (0.81, 0.59 and 0.29 for the second, third, and fourth groups of tests). Overall, this data indicates that most participants initially relied on the same type of information to perform the verification and naming tasks, but that the informational basis of Naming decisions changed over the course of training, while the informational basis of Verification decisions remained roughly the same.[2]

2.3. Overview of expertise training

Most participants in our expertise training eventually became as good at identifying Greebles at the individual level as they were at identifying the Greebles at the gender level. Gauthier and Tarr [4] took this pattern of performance as their criterion for demonstrating perceptual expertise with the Greebles, based on the fact that the same pattern of results was found by Tanaka and Taylor [7] for dog and bird experts. In Gauthier and Tarr's procedure, participants practiced identifying Greebles over and over through the same verification test that served as the criterion for expertise. In contrast, the present procedure effectively substituted trials on the naming task for trials Verification test, as these tests were alternated throughout training. Furthermore, our procedure required participants to learn twice as many Greebles (20 vs. 10) as did Gauthier and Tarr's procedure [4], which should have made it much more difficult to individuate the Greebles whose names had to be verified on the verification task. The fact that most of our participants reached the verification test criterion after roughly the same amount of training as the participants in Gauthier and Tarr's study indicates that this criterion may reflect something general about perceiving Greebles, and is not purely task-specific.

The results of Section 2 also raise several questions about the factors infiuencing object recognition expertise training. For example, participants learned to name Greebles that they were experienced with through NIL trials more rapidly than when they did not have this prior experience. One course for future study is to investigate how much of this effect was due to information participants acquired about Greebles in general and how much due to information about those specific Greebles. Another issue involves the relationship between the verification and naming tests. Conceptually, these two tasks seem very similar, and they are often treated as more or less equivalent in the object recognition literature. The present results suggest that while participants initially used similar types of information on both tests, the informational bases for the two tasks diverged over the course of training. Additional research is needed to determine whether this finding represents a general property of perceptual expertise or a specialized strategy developed by our participants in response to the specific demands of our training procedure and stimulus set.

3. Learning new sets of Greebles

Intuitively, the most obvious way to assess expertise with a perceptual category is to determine how well experts learn new exemplars of the category. After their training had been completed, our experts learned four sets of six new Greebles. Sets A and B (Fig. 6) were presented immediately following the final training session,[3] while the other two sets of test Greebles, C and D, (Fig. 6) were learned in a second test session that was separated by 8–13 weeks from the last training session (only ten of the 12 experts participated in the second test session). If our trained participants were truly Greeble experts, we expected them to be able to learn all new sets of Greebles more easily than control groups of novice participants who received no prior training with the Greebles (separate groups of 12 novices each learned the first and second pairs of test sets).

In designing the four test sets, we were interested in what information participants relied upon when learning new Greebles. Sets A and B were taken from the same pool of Greebles as the training objects, each of the two sets including Greebles from a single gender and three Greebles from each of two families. As in the training set, then, these test sets included Greebles with fairly heterogeneous collections of appendage parts,

[2] The same pattern of Naming and Verification diverging over the course of expertise of training and of Naming changing more than Verification was recently replicated in an ongoing training study by the first author. This occurred even though this second cohort of Greeble experts were trained on the individual and family levels rather than on the individual and gender levels as in the current study, suggesting that this result does not depend on specific factors of the training procedure.

[3] Unfortunately, we mistakenly used one of the Greebles from the original training set as a member of Test Set B. In response to debriefing questions, six of the 12 experts spontaneously mentioned this correspondence, but all of them said only that the Greeble on the test set 'looked like' or was 'similar to' the training Greeble, and several of these experts also mentioned at the same time that other test Greebles were similar to training Greebles. Four of the six experts who mentioned noticing the correspondence claimed that it hurt their test performance, because they had to associate the Greeble with a new name. An examination of mean performance for each Greeble confirmed this intuition: Experts' performance on this Greeble was less accurate than all of the other 11. Greebles from the first two test sets and slower than all but two of the other Greebles. Therefore, we left the repeated Greeble in our analyses, since doing so could only weaken our hypothesis that experts were better at learning new Greebles than were novices.

Fig. 6. Four sets of Greebles learned by both experts and novices during test sessions.

while body shape information, which defines family membership, was relatively uninformative. Sets C and D were both constructed with more homogeneous parts than any randomly selected set of six training Greebles. However, in designing set C we made the body shape information maximally diagnostic (that is, each Greeble has a different body shape) while in set D all Greebles shared exactly the same body shape.

The procedure for teaching participants names for test sets involved the same tasks as were used in expert-training sessions, and is summarized in Table 2. Note that all participants performed the naming test at least three times on each set of test Greebles, but if a perfect score was not obtained on the third test, the participant was required to repeat the naming with feedback and naming tasks until this criterion was met (or until they had completed 13 naming tasks).

One measure of learning facility is how quickly participants passed the criterion of a perfect score on one test. As seen in Table 3, the mean number of tests to reach criterion was smaller for experts than for novices on all four test sets. The difference between experts and novices was not significant for any one test set alone, but was significant when all four sets were considered together, $t(90) = 2.00$.[4]

Table 2
Training procedure for test sets

Trials	Task
1	Individual inspect (six Greebles at once)
6	Naming with response
6	Naming with feedback
12	Naming
6	Naming with feedback
12	Naming
6	Naming with feedback[a]
12	Naming[a]
61[a]	Total number of trials

[a] If performance on the third naming task was not perfect, participants performed the naming with feedback and naming tasks again, and continued to cycle through these tasks until a perfect score was obtained on the naming task, or until 13 cycles had elapsed.

Table 3
Tests to reach criterion for test sets

Set	*N*		Mean tests to criterion	
	Experts	Novices	Experts	Novices
A	12	12	2.0 (1–6)	3.8 (1–12)
B	12	12	2.3 (1–4)	3.7 (1–9)
C	10	12	3.2 (1–6)	4.6 (1–14[a])
D	10	12	5.8 (2–10)	6.0 (1–14[a])
Mean			3.2	4.5

[a] Some novices (one on the heterogeneous set and two on the homogeneous set) failed to pass the criterion in the 13 tests we allowed. In calculating the means in this table, we conservatively assumed that these participants would have passed on the next (14th) test.

[4] For this and subsequent analyses in this Section, we treated each participant's performance on each test set as an independent observation, because some expert participants learned all four sets, some experts learned only two of the sets, and all novices learned only two sets each.

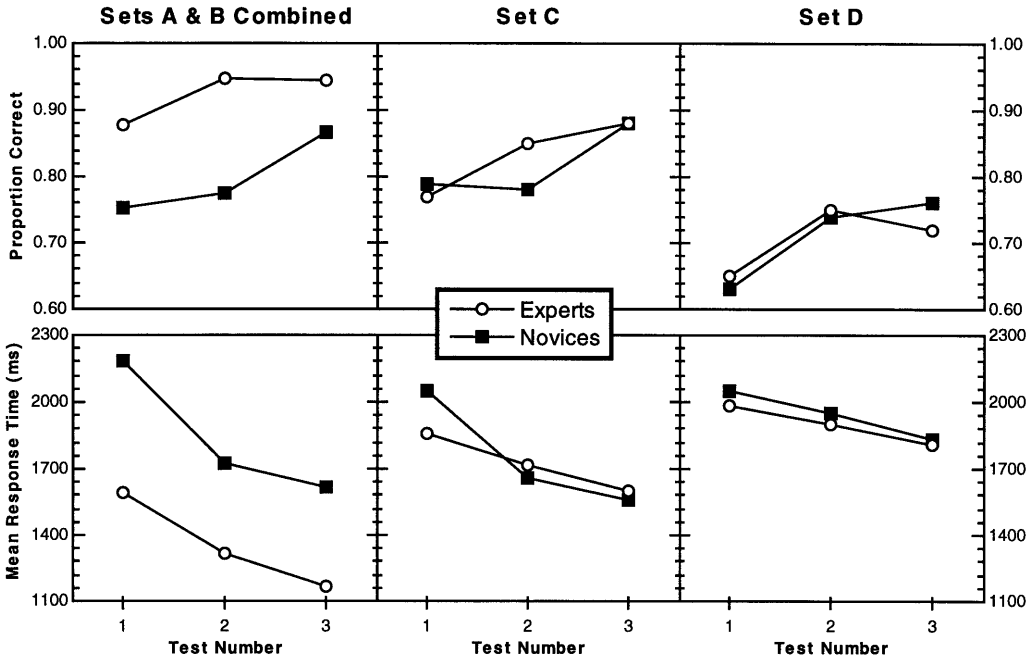

Fig. 7. Performance by experts and novices on the first three tests when learning new sets of Greebles.

Since participants were required to perform at least three tests on each set of Greebles (even if they passed the criterion on the first or second test), we were also able to compare experts' and novices' performance on each of these three tests. ANOVAs including the factors of expertise, test set, and test number revealed significant main effects of all three variables on both accuracy and response time (expertise: $F(1, 252) = 7.19$ for accuracy and $F(1, 252) = 14.8$ for RT; test set: $F(3, 252) = 8.89$ and $F(3, 252) = 4.78$; test number: $F(2, 252) = 5.16$ and $F(2, 252) = 12.9$). The interaction of expertise and test set was marginally significant for accuracy and significant for RT, $F(3, 252) = 2.15$ and $F(3, 252) = 4.36$. No other interactions approached significance. Inspection of the data revealed that patterns of effects were very similar for test sets A and B, but were somewhat different for set C and set D. Therefore, Fig. 7 shows results for the first two sets combined, and for the second two tests separately.

For the first two test sets, experts were considerably more accurate and faster than novices at recognizing test Greebles on each of the first three tests (Fig. 7). Note that experts achieved 88% accuracy on the very first test, after seeing the test Greebles only three times each, and were nearly perfect (95% accuracy) on the second test, after seeing the Greebles six times each. In contrast, novices were only accurate on 77% of trials by the second test.

Test Set D was apparently more difficult to learn than set C for both novices and experts, suggesting that all participants relied on body shape information when it was diagnostic for individual discrimination. Experts did not show a marked advantage over novices for either of these two test sets. This could be due to either or both of two factors. First, these test sets were learned some 8–13 weeks after the conclusion of experts' training, so it is possible that participants 'lost' their expertise during this time. Second (and, we feel, more likely), the Greebles in these test sets contained parts that were more homogeneous than those in the original training set or in test sets A and B. Thus discriminating two Greebles from test set D may have been as difficult for Greeble experts as discriminating twin brothers would be for face experts.

In other words, experts may not have been especially proficient at extracting the information required to discriminate Greebles in the second pair of test sets. Rather than using their specialized knowledge of how to discriminate Greebles such as the ones in the training set, where body shape was relatively uninformative and appendage parts were highly diagnostic, experts presumably had to revert to the same strategies used by novices. A correlational analysis of performance on the Greebles in the second and first pairs of test sets (collapsed across the first three tests on each set) provides some support for this hypothesis. For the second pair of sets, the correlation of expert and novice accuracy over the 12 Greebles was 0.71. For the first pair of sets, this correlation was 0.56 for all 12 Greebles, but this number is inflated by one Greeble on which experts

and novices were accurate 100 and 98% of the time, without this Greeble, the correlation dropped to 0.26. Thus for the first pair of tests, experts and novices seem to have used different sources of information to learn the Greebles (since different Greebles were difficult and easy to learn for the two participant groups), whereas for the second pair of tests, experts and novices used the same sources of information (however, see Section 6 for evidence that experts nonetheless processed the Greebles in set C differently than novices).

To summarize, the results of Section 3 indicate that our experts did 'learn how to learn' during the training procedure, in that: (a) they reached a perfect-accuracy criterion faster than novices for all four Greeble test sets; and (b) were faster and more accurate than novices on the first three training tests for test sets A and B. However, the expertise derived from our training procedure appeared not to have transfered well to the second two test sets, in which Greebles were relatively harder to distinguish on the basis of individual parts. This may indicate that one aspect of the expertise resulting from our training procedure is knowledge of the specific ways in which Greebles in the training set differ. That is, experts may have acquired a psychological similarity space [8] capturing the variation among Greebles in the training set, and became very good at extracting the diagnostic information about any one Greeble that differentiated it from other members of the set. Since Greebles in the first pair of test sets were drawn from the same pool as the training set, experts were better at learning to distinguish these Greebles from each other than were novices, who did not have an *a priori* representation of the similarity space for the set. Greebles in the second pair of test sets effectively formed different similarity spaces from the space of training Greebles, and therefore were as difficult to learn for experts as for novices.

The present findings are consistent with other examples of expertise failing to generalize to subcategories of objects. Diamond and Carey [6] found the inversion effect (a large advantage in performance when stimuli are studied upright rather than inverted) in dog show judges to be much larger for breeds in their domain of expertise as compared to other breeds. Myles-Worsley et al. [9] report that radiological expertise with X-ray films is specific to clinically significant deviations. Moreover, we all experience from time to time the limits of our face expertise, when having to discriminate twins or people from a less familiar race—indeed, Rhodes et al. [10] found that the inversion effect was larger for faces of the participants' own race than for different-race faces.

4. Sensitivity to configural information

The results of Section 3 suggest that one of the ways in which experts and novices differ is in their knowledge of what kinds of information are helpful in distinguishing different Greebles from one another. Experts and novices may also differ in the way they process distinguishing information about Greebles. Specifically, prior studies indicate that experts process information 'configurally'. Although the term has been used somewhat loosely in the past, we define configural processing as the ability to take into account the precise relations between different parts of objects as well as the parts themselves.

Much of the evidence that expertise leads to configural processing is based on the inversion effect, and is thus indirect: we know that inversion disrupts recognition of objects for which people are experts (faces for all humans [11–13] and dogs and birds for dog and bird experts [6]), and inversion is thought to disrupt the use of configural cues more than the use of featural cues [14–16]. Tanaka and his colleagues [17,18] developed a more direct way of testing whether the relationships between parts are important for recognition of objects in a given class. They had participants recognize parts of faces or control stimuli (scrambled faces, inverted faces, or houses) in a forced-choice procedure where the parts were presented: (1) in their original configurations (Jim's nose in Jim's original face); (2) in a transformed configuration (Jim's nose in Jim's face with his eyes slightly moved apart); or (3) in isolation (Jim's nose alone). If participants used independent part representations to perform the task, there should have been no difference between the three conditions. Instead, results showed that parts of upright faces (but not of control stimuli) were better recognized in the original configuration than in the transformed configuration or in isolation. Crucially, the fact that moving the eyes slightly apart impaired the recognition of the nose and mouth provided strong evidence that participants could not ignore relations between face parts, even when told to do so.

Gauthier and Tarr [4] showed that the old/new configuration advantage (better part-identification in the original than in the transformed configuration) can be obtained for Greebles with experts but not novices, suggesting that the crucial dimension in the Tanaka studies was not stimulus category but rather participant expertise. Gauthier and Tarr [4] as well as Tanaka et al. [19] found that both novices and experts could display a whole/part advantage with stimuli such as Greebles, cars or cells. Tanaka and Gauthier [2] interpreted this as evidence that although novices may sometimes rely on first-order relational properties (e.g. the quiff is located between the boges), only experts seem to rely on second-order relational properties (e.g. the angle of

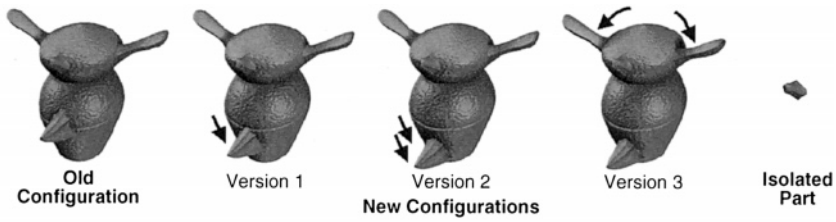

Fig. 8. Stimulus conditions for the configural test.

the boges). Section 4 investigates whether our experts would also demonstrate such configural sensitivity when attempting to recognize Greeble parts. In Gauthier and Tarr [4], the change in the transformed configuration was so subtle that most novices did not notice it. Here, we used three increasingly obvious configural transformations, the most radical of which was readily perceivable by novices. This procedure allowed us to test whether the performance costs associated with transformed configurations is related to the amount of configural change, and whether novices would also show a configural effect when they explicitly noticed the change in configuration.

4.1. Method

4.1.1. Participants

The 12 experts trained in Section 2, as well as a second group of 11 novices (who had no prior exposure to the Greebles), served as participants. An additional novice's data had to be dropped because he or she used the wrong keys in the test.

4.1.2. Materials

Six Greebles (test set A, Fig. 6) that were not used in the training, and were thus equally unfamiliar to novices and experts, served as stimuli. For each Greeble, four target versions were generated (Fig. 8): in Version 0 (old configuration), all parts were in their original positions; in Version 1, the dunth (bottom part) was moved down slightly; in Version 2, the dunth was moved down even more; and in Version 3, the boges (top parts) were moved 15° around the vertical axis towards the front. Distractors were also created for each version of each Greeble; distractors were identical to targets, except that one of the three parts (boges, quiff, or dunth) was replaced by the corresponding part from a different Greeble from the test set. Finally, images of the three parts of each Greeble in isolation were also created. Each isolated part served once as a target and once as a distractor.

4.1.3. Procedure

Both novices and experts were first taught to name the six test Greebles, as described in Section 3. They then received instructions for the Configural Test. Each trial consisted of a 1000 ms blank screen, a 250 ms fixation cross, a prompt, shown for 2000 ms, specifying one part of a particular Greeble (e.g. 'SOSFA'S BOGES'), and finally two stimuli shown side-by-side, which stayed on the screen until participants responded. One stimulus was the target, and showed the specified part (SOSFA's boges) on Version 0, 1, 2, or 3 of the specified Greeble (SOSFA). The other stimulus was a distractor, and showed a different part (e.g. FERZU's boges) on the specified Greeble (SOSFA). Participants were to select whether the right or left image contained the designated part by pressing one of two keys. Each individual part was the target in four trials: once each embedded in Versions 0, 1, 2, and 3 of its Greeble.

Following these 72 trials (6 Greebles $\times$ 3 parts $\times$ 4 versions), participants performed the same task on the isolated Greeble parts. The 18 isolated-part trials (6 Greebles $\times$ 3 parts) were separated from the configural-change trials because Gauthier and Tarr [4] found that isolated parts were recognized considerably faster than parts in the context of whole Greebles, and we hoped that by separating the two types of trials, we might reduce some of the variance in each of the two tasks.

4.1.4. Design

Dependent measures were response time and accuracy. Group (novice or expert) was manipulated between participants, while Part (boges, quiff, or dunth) and stimulus condition (old configuration, new configurations 1, 2, or 3, or isolated part) were within-participant variables.

4.2. Results

A preliminary analysis including only the three new configurations (versions 1, 2, and 3) showed no main effect of Stimulus Condition, and this factor did not significantly interact with either Group or Part (all F's < 1). Therefore, these conditions were collapsed in subsequent analyses, in which the Stimulus Condition factor included three levels: old configuration, new configuration, and isolated part.

An ANOVA on response times revealed significant main effects of all three factors: Experts were faster than novices (2352 and 3345 ms, respectively), $F(1, 21) = 4.45$; boges (2130 ms) were more quickly recog-

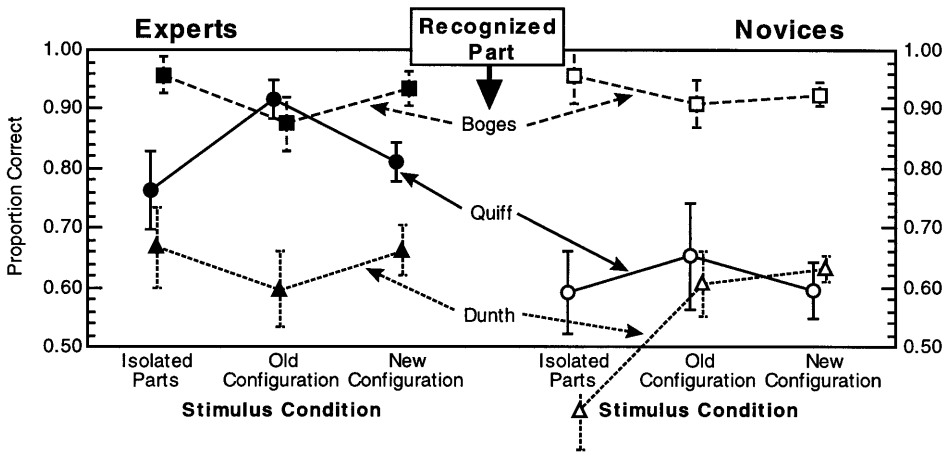

Fig. 9. Accuracy rates for experts and novices on the configural test. Error bars represent standard errors of the means for each individual condition.

nized than quiffs (3425 ms) which were more quickly recognized than dunths (3711 ms), $F(2, 42) = 19.1$; and the isolated part condition (2446 ms) was faster than either of the other two conditions (2985 and 2993 ms for old and new configurations, respectively), $F(2, 42) = 7.31$. However, no interactions approached significance for response times.

Accuracy rates for each group × part × stimulus condition combination are shown in Fig. 9. An ANOVA on this data revealed significant main effects of Group, $F(1, 21) = 8.02$, and Part, $F(2, 42) = 58.34$, significant two-way interactions of Group × Part, $F(2, 42) = 6.59$, and Part × Stimulus Condition, $F(4, 84) = 2.86$, and a marginally significant three-way interaction, $F(4, 84) = 2.32$, $P = 0.064$. To further investigate the latter effect, we performed separate ANOVAs on experts' and novices' data, and computed post-hoc LSD tests to determine when the old configuration condition was significantly different from the new configuration or isolated part conditions. This analysis indicated that for experts, quiffs were significantly easier to recognize in their old configurations than as isolated parts, $P < 0.01$, or in their new configurations, $P = 0.059$. For novices, dunths were easier to recognize in their old configurations than as isolated parts, $P < 0.05$, although since performance in the isolated part condition was below chance, the significance of this result should be interpreted with caution.

4.3. Discussion

In a part recognition task that was considerably different than all the tasks used during expertise training, our experts performed more accurately and faster than novice participants, indicating again that their expertise was quite generalizable across tasks and Greebles. A second result was that participants were more accurate in the old configuration condition than in the isolated parts condition. However, as in other studies [4,19], this was true for both experts and novices; furthermore, participants were also substantially faster in the isolated parts condition, indicating the possibility of a speed-accuracy trade-off [4]. Therefore, the part/whole advantage should not be taken as an expertise effect here.

On the other hand, a significant old/new configuration advantage was found for experts but not novices, and is not subject to speed-accuracy tradeoffs. This expert-novice difference was obtained despite the fact that configuration changes were readily perceivable by novices, unlike in Gauthier and Tarr's study [4]. In the present study, the old/new advantage was only found for quiffs, not for dunths or boges. It is interesting that the only part not moved in the new configuration condition was the one for which we found the predicted old/new configuration advantage. However, previous studies [4,18] found the whole/part advantage and old/new configuration advantage with all tested parts, not just a single part.

While the present results could be interpreted as a failure to replicate Gauthier and Tarr's findings [4], several differences between the training and test methodologies of the two studies could be responsible for the discrepant results. One potentially problematic (in hindsight) aspect of the present experiment could be the particular set of Greebles chosen for testing. As is obvious in Fig. 6, the boges of these Greebles (set A) were all highly distinctive; compare these Greebles to set B, where no one part is as uniformly diagnostic. Furthermore, participants would have been alerted to the diagnosticity of the boges at the beginning of the learning procedure, when the six Greebles were shown on the screen in a group (neither Gauthier and Tarr, nor Tanaka and Sengco, ever showed more than one stimulus at a time during learning [4,18]). In fact, most experts reported on a post-test questionnaire that they

focused on the boges to learn these Greebles, and both experts and novices performed practically at ceiling in recognizing these parts on the test. Furthermore, the dunths, which were the parts farthest from the boges and thus presumably farthest from participants' focus of attention during training, were very poorly recognized by both novices and experts. The quiffs, which were spatially close to the boges but were not particularly diagnostic for recognition, were much better recognized by experts than by novices, and were the part for which a significant old/new configuration difference was found for experts (but not for novices).

Although post-hoc, this account is broadly consistent with the view that object recognition is based neither on undifferentiated images nor on a fixed vocabulary of features/parts—rather, objects are represented in terms of functional features that are acquired based on the training context and can be modified by experience [20,21]. Thus all participants may have focused on the boges to distinguish between the test Greebles, but experts used the larger 'head' region, including both the boges and the quiff.

One caveat may be in order: Ashby and Maddox [22] have shown that it is logically possible for subjects to demonstrate 'redundancy gains' (i.e. facilitation from a to-be-ignored dimension when attention is directed to a second dimension) in the absence of configural processing. More specifically, their results indicate that tasks such as the part recognition test and composite test (Section 5) used here cannot provide definitive evidence that participants are processing features configurally or holistically; rather, participants could simply be using an optimal decision strategy. However, Ashby and Maddox [23] found that participants were not using an optimal criterion after 100 trials of practice categorizing simple stimuli (rectangles). It is thus unlikely that our experts could have reached an optimal decision strategy in the configural test used here given the limited number of trials (90) and the complexity of the stimuli.

5. Recognition of composite greebles

Another experimental task on which differential effects have been shown for faces and non-face objects is the 'composite task'. Young et al. [24] had participants identify the parts of composites made out of the top and bottom halves of two different faces, and found that part-recognition was slower when the two halves of the composite were aligned (effectively forming a new face) than when misaligned. Intuitively, the composite effect seems to reflect the same kind of configural processing revealed by the old/new configuration advantage [18] and the inversion effect [13]. However, Carey and Diamond [5], on the basis of developmental evidence, suggested that the composite effect may reflect a more general mechanism that can be termed holistic processing. We define this type of processing as the ability or tendency to consider all parts of an object simultaneously, regardless of the exact configuration of the parts.[5] Carey and Diamond's hypothesis is that when halves of two different faces are aligned, one cannot help but consider the entire stimulus as a single object, so the half that is not supposed to be recognized has an interfering effect. With misaligned composites, on the other hand, we can consider the two halves separately, ignoring the distracting information from the half that is not to be recognized.

We initially probed for differences between our experts and novices in a very difficult version of the composite test (employing stimuli composed of portions of three different Greebles each) using Greebles from test set B, but found no meaningful differences between the two participant groups or between aligned and misaligned Greebles. This null result was preceded by two other failed attempts to find a composite effect with Greebles (unpublished pilot experiments performed on the experts from Gauthier and Tarr's study [4]). In the present experiment, we attempted to replicate as closely as possible the conditions that have proved most amenable to a composite effect in the past. First, stimuli were composed of the top half of one Greeble combined with the bottom half of another, just as the top and bottom half of faces were used in previous studies. Second, we used the 20 Greebles on which experts were trained in Section 2, since there is some evidence that the composite effect may be easier to obtain with famous faces ([24]; also see [25]). These Greebles were familiar only to experts (and the test procedure requires that participants know the names of the Greebles), so we were not able to test novices. Finally, because when famous faces are used it is the faces that are assumed to be famous and not the particular pictures, we used mirror-images of the 20 familiar Greebles to reduce image-based similarity.

In addition to testing composites made up of two different Greebles, we also included trials in which the two halves of famous Greebles were presented either aligned or misaligned. Our prediction was that when the two halves of a stimulus were from different Greebles, experts would be faster when the halves were misaligned than when they were aligned (this is the traditional composite effect), whereas when the two halves were from the same Greeble, experts would be faster for aligned than for misaligned stimuli. For the composite stimuli, the two Greebles used could either

[5] Note that the terms 'configural' and 'holistic' are not meant to imply a pixel-like representation based on linear coordinates. Rather, we assume only that local features, compositional or otherwise, areinterdependent in a manner that leads to sensitivity to configural changes.

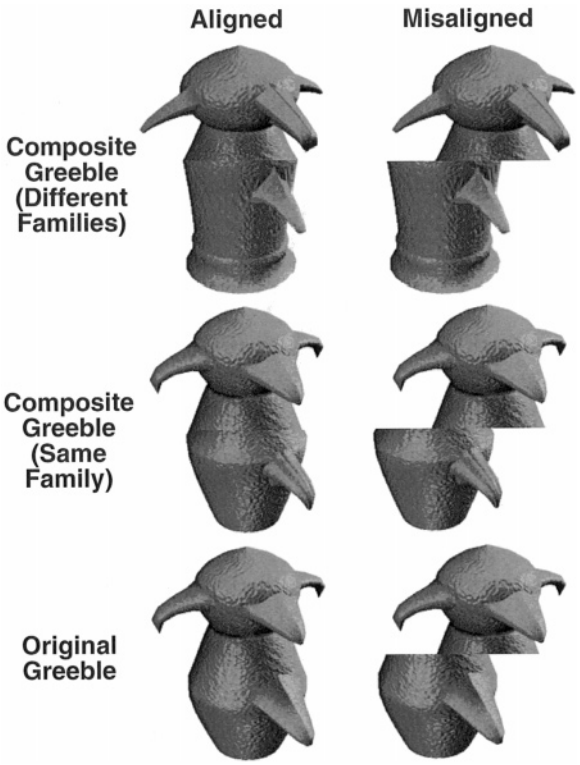

Fig. 10. Stimuli for the composite test.

be members of the same family (i.e. share the same body shape) or members of different families (Fig. 10). Assuming that the composite effect is caused by participants 'fusing' the two halves of the stimuli, we might expect the effect to be stronger for same-family composites, since they should be easier to fuse.

5.1. Method

5.1.1. Participants

Participants were the ten experts from Section 2 who returned for the second test session, 8–13 weeks after the conclusion of training.

5.1.2. Materials

Stimuli were constructed by splitting in half and recombining the 20 Greebles for which participants learned names in the training. For each Greeble, four versions were tested (Fig. 10): two halves of different Greebles (composites) aligned or misaligned, and two halves of the same Greeble (originals) aligned or misaligned. Composites could either combine two Greebles from the same family (body type), or two Greebles from different families. All images were mirror-reversed, to disrupt 'template-matching' processes that experts might otherwise have used on these highly-overlearned images.

5.1.3. Procedure

Prior to the test, participants received two blocks of the naming with feedback task, to remind them of the Greebles' names (performance on these tasks is described in Section 2). In the composite test, participants saw each stimulus twice, once preceded by the prompt 'QUIFF' and once preceded by the prompt 'DUNTH'. They were required to press the key corresponding to the first letter of the name of the Greeble from which the prompted part came from. That is, if the prompt was 'QUIFF' and the top half of the stimulus was from the Greeble 'VALI', the correct response was 'V'. To prevent participants from focusing solely on the prompted region, each stimulus appeared at a random position on the screen.

5.1.4. Design

Dependent measures were response time and accuracy; independent variables were Version (original, composite-same family, or composite-different families) and Alignment (aligned or misaligned), both within participant. Because of an experimenter error, the part judged on each trial (quiff or dunth) was not recorded, so this variable could not be analyzed.

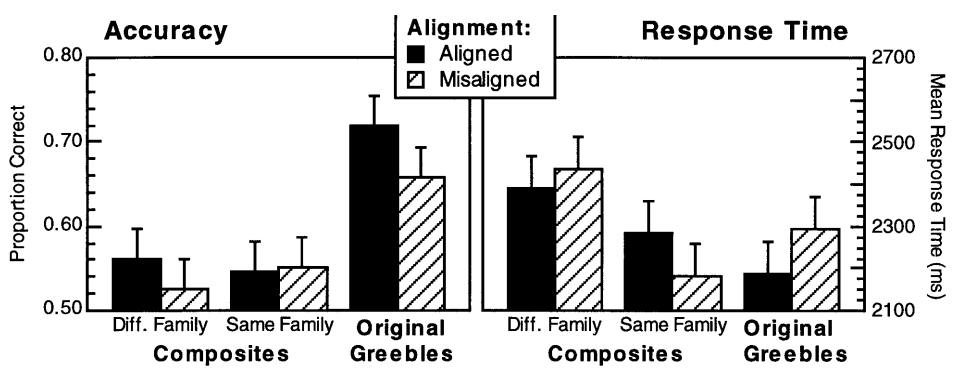

Fig. 11. Performance on the composite test. Error bars represent standard errors of the mean, calculated based on the interaction term of the analysis of variance [59].

5.2. Results

Mean accuracy rates and response times are displayed in Fig. 11. Note that chance performance in this task would be 5% correct, as there were 20 possible responses. The ANOVA on accuracy revealed a significant main effect of Version, $F(2, 18) = 9.67$, and post-hoc Scheffe tests revealed that the original condition was significantly easier than either of the two composite versions, which did not differ significantly from each other. Neither the main effect of Alignment nor the interaction of the two factors was significant. For response times, neither the main effects nor their interaction was significant. However, we also performed an analysis restricted to the response times from the original and same-family composite conditions because these conditions were most likely to lead to a 'fusion' of to the two parts of each stimulus. This analysis revealed a marginally-significant interaction of version and alignment, $F(1, 9) = 4.63$, $P = 0.06$: for original (non-composite) Greebles, participants were faster on aligned than misaligned versions; while for same-family composite Greebles, participants were faster on misaligned than aligned versions.

5.3. Discussion

Participants identified parts of composite Greebles faster when they were misaligned than aligned, but identified parts of non-composite (original) Greebles faster when they were aligned than misaligned. This experiment thus provides the first evidence of a composite effect for non-face objects. However, this conclusion depends on dismissing results from the different-family composites, as participants did not show a composite effect for these items. In our view, these composites may have been processed similarly to misaligned, same-family composites: since experts were very familiar with the five Greeble body shapes, the new body shapes that emerged from different-family composites may have aided experts in processing the two halves separately.

In an ongoing study (Gauthier and Tarr, in preparation), we obtained a conceptual replication of this result using a composite test on famous faces. In this experiment, we showed stimuli that were made up of two halves of faces that were either the same or different genders. For example, the top half of Tom Cruise's face could be paired with the bottom half of Mel Gibson's face or the bottom half of Princess Diana's face. The composite effect (faster response times to identify face parts in misaligned than aligned composites) was over twice as large for same-gender composites as for different-gender composites. Thus gender information for faces, like body-shape information for Greebles, seems to provide a cue to the visual system about when and when not to 'fuse' two halves of a composite stimulus.

These results probably reflect a mix of configural processing and more general holistic processing. Configural processing (consideration of the relations between parts as well as the parts themselves) is indicated by participants' faster and more accurate responses to original (non-composite) Greebles when aligned than when misaligned. That is, when aligned, parts of the original Greebles were in the correct configuration, and participants were able to use this configural information to their advantage. Holistic processing (consideration of all parts together, regardless of whether they are in the proper configuration or not) is indicated by the fact that participants were much more accurate on original than composite Greebles (regardless of alignment), showing that having all parts present, even when in the wrong configuration, is beneficial to recognition. Note that whereas the original Young et al. task is thought to reflect holistic processing, they never tested recognition in original faces with aligned vs. misaligned parts. It is this condition that allows us to also test for both configural and holistic processing in Part 5. Moreover, for same-family composites (and not for originals), response times were faster for misaligned than for aligned stimuli. Following Carey and Diamond [5], this last finding indicates that parts of aligned composites were processed holistically, forcing participants to consider parts that were not relevant to the task but that nonetheless had an interfering effect on performance.

Since novices could not be tested with the present procedure, the results of Section 5 are not informative as to whether the composite effect is due to expertise *per se*. Instead, the effect could be attributable to the complex nature of the stimuli themselves (i.e. the fact that all Greebles share the same basic configuration), a conclusion drawn by Gauthier and Tarr [4] for the whole/part advantage (see also [2]). Carey and Diamond [5] have shown that the holistic processing of the kind tested by the original Young et al. paradigm (our 'composite' conditions) does not increase after 6 years of age. It remains possible that experience with faces before that age is sufficient to mediate this effect. The resolution of this issue awaits further study. Nevertheless, the fact that we obtained the composite effect with Greebles eliminates it from the ever-shrinking list of effects that can be taken as evidence for a face-specific recognition module.

6. Recognition of transformed Greeble images

The experiments reported in Sections 4 and 5 evaluated the nature of the abilities acquired by experts in processing Greebles. Another important issue in perceptual expertise is the conditions under which experts can and cannot perform better than novices; in other

words, the conditions under which expertise generalizes. At one extreme, extended practice with a particular picture of a specific exemplar should make an observer highly expert at processing that one image. At the other extreme, expertise almost never generalizes across basic-level categories—for instance, becoming a dog expert does not make one a bird expert. In-between these two extremes lie the conditions under which expertise transfers from known objects of a class to new members of the same class.

In searching for ways to evaluate the generalizability of Greeble expertise, we again turned to the face recognition literature. A classic 'face-specific effect' is the inversion effect [26,12,13]: Faces encoded upside-down (inverted) are more difficult to recognize than upright faces, and this difference in difficulty is greater for faces than for other objects. Diamond and Carey [6] suggested that we may not be able to encode inverted faces using our expert abilities developed on upright faces, although once a face is encoded at the upright, we may be able to recognize it in other orientations via normal object recognition processes.

It could also be that for a class of objects encountered primarily in a single orientation, experts develop a strong advantage for this 'canonical' orientation over all others. Supporting this interpretation, Tarr and Pinker [27] showed that subordinate-level recognition of novel object classes does show such viewpoint specificity, and Tarr and Gauthier [28,29] further found that these orientation effects generalize across members of a class. That is, if one exemplar of a class is easiest to recognize in a particular viewpoint, other visually similar members of the class may also be easy to recognize at that viewpoint. Moses et al. [30] provide specific evidence for such class-based generalization in faces. These findings all suggest that expert recognition should be strongly viewpoint-dependent even for relatively unfamiliar exemplars, provided that objects in the domain of expertise are generally experienced from a single viewpoint. We tested this prediction by having expert and novice participants name upright and misoriented Greebles that had been learned in an upright orientation. Note that this is not strictly speaking a test of the 'inversion effect', since Greebles will be encoded at the upright and tested in other orientations. Indeed, Carey [31] stated that "…the difficulty [in the inversion effect] is in forming an adequate representation of an inverted face, not in coping with a mismatch of orientation between test and recognition".

A second transformation that appears to disrupt face recognition is brightness reversal: pictures of faces presented in a photographic negative, thus inverting the brightness level of each pixel in the pictures, are more poorly recognized than normal pictures [32–35]. If the brightness-reversal effect is mediated by expertise, experts but not novices should be impaired at recognizing brightness-reversed Greebles. This prediction is also tested here.

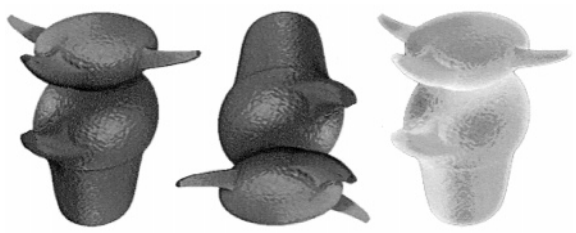

Fig. 12. Stimuli for the rotation and brightness-reversal tests: normal (left), 180°-rotated (middle), and brightness-reversed (right) Greebles.

The experiments reported in this Section are concerned with the transfer conditions of expertise. Experts are, by definition, better than novices at recognizing objects of a given class under some conditions. However, the experiments reported here test the paradoxical prediction that under conditions that differ from those employed during training, experts may actually be worse, or at least no better, than novices. More specifically, we predicted that novices would be able to adapt to new stimulus-presentation conditions (misorientation and brightness reversal) more readily than experts.

6.1. Method

6.1.1. Participants

For the Rotation Test, the 12 experts from Section 2 and a separate group of 12 undergraduates from Oberlin College served as participants. For the brightness-reversal test, ten of the 12 experts and a third group of 12 undergraduates from Brown University participated.

6.1.2. Materials and procedure

The rotation test was performed on test set B. Each trial consisted of a 1000 ms blank screen, a 250 ms fixation cross, and a Greeble in its familiar upright orientation or rotated 60, 120, or 180° in the picture plane (Fig. 12). Participants were required to identify the Greeble by pressing the key corresponding to the first letter of its name; the Greeble remained on the screen until participants responded. One half of the rotations were clockwise and the other half counterclockwise in the picture plane. All six test Greebles were presented in every orientation twice, once each in two blocks of 24 trials. The order of trials in each block was randomized.

The brightness-reversal test was performed twice, once on test set C and once on test set D. Although two of the novices failed to reach the criterion in learning one or both of the test sets, they were still significantly above chance on the brightness-reversal test, so their data was included in the analyses. Each trial again consisted of a 1000 ms blank screen and a 250 ms fixation cross, followed by a test Greeble, which re-

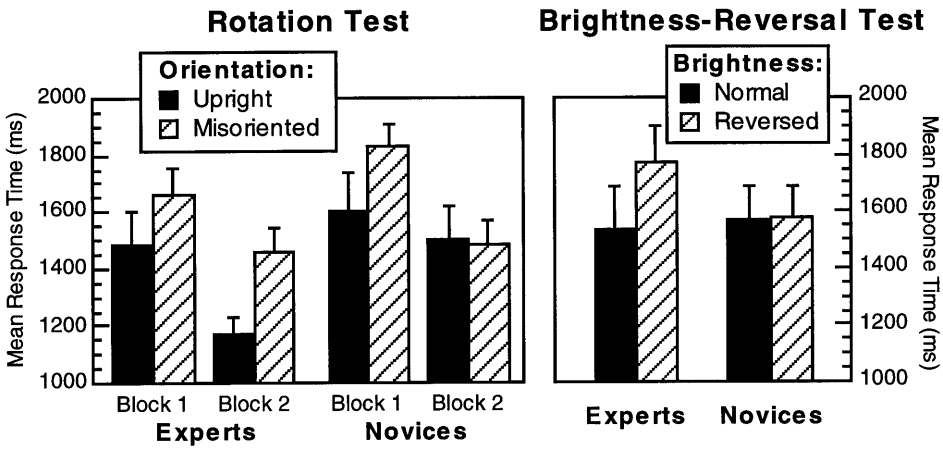

Fig. 13. Mean response times for experts and novices on the rotation test (left) and brightness-reversal test (right). Error bars represent standard errors of the means for each individual condition.

mained on the screen until the participant responded by pressing the key corresponding to the Greeble's name. On one half of the trials, the Greeble was displayed in its normal color scheme, while in the other half, the brightness of each pixel in the test image was inverted (Fig. 12). All test Greebles were presented twice each in their normal and inverted forms.

6.1.3. Design

Accuracy and response times served as the dependent variables in both tests. Relevant independent variables were group (expert or novice), and block (1 or 2), and rotation angle (0, 60, 120, 180) for the rotation test, and group, test set (heterogeneous or homogeneous), and brightness (normal or inverted) for the brightness-reversal test. Group was a between-participants variable, while the other factors were all manipulated within-participants. Since there were relatively few observations per participant per cell, and relatively few participants, we were concerned that outlier response times would bias our results. Therefore, we excluded response times greater than 5000 ms from analysis, eliminating the slowest 3.7 and 4.9% of the correct responses from the rotation and brightness-reversal tests, respectively.

6.2. Results: rotation test

Preliminary ANOVAs including the factor of rotation angle (0, 60, 120, 180°) revealed a significant main effect on both response times and accuracy, but post-hoc tests indicated that differences between the misoriented conditions were miniscule and not significant (mean response times of 1621, 1621 and 1611 ms for 60, 120 and 180° orientations, respectively). Therefore, these rotation angles were combined in subsequent analyses, which included an orientation factor (upright or misoriented).

Compared to novices, experts were numerically, but not significantly, faster, $F(1, 22) = 1.63$, $P = 0.22$ and significantly more accurate, $F(1,22) = 5.96$ overall. Both experts and novices were faster and more accurate on the second than on the first block of trials; this effect was significant for RT, $F(1,22) = 31.2$, and marginally significant for accuracy, $F(1,22) = 3.73$, $P = 0.07$. The main effect of Orientation was significant for both RT and accuracy, $F(1,22) = 16.7$ and $F(1,22) = 7.74$ respectively, indicating that upright Greebles were recognized considerably easier than misoriented ones. Two interactions were also significant: block × orientation for accuracy, $F(1,22) = 6.15$, and group × block × orientation for RT, $F(1,22) = 8.87$. The former interaction reflects the fact that for accuracy, the difference between upright and misoriented Greebles was smaller for the second block of trials than the first. More interestingly, the three-way interaction of group, block, and orientation indicates that for RT, this effect occurred for novices, but not for experts. This interaction is shown in Fig. 13: Novices were 231 ms faster at recognizing upright compared to misoriented Greebles on the first block of trials, but this difference disappeared (-13 ms) on the second block. For experts, on the other hand, the difference between upright and misoriented Greebles was actually larger in the second than in the first block (block 1: 180 ms; block 2: 274 ms).

6.3. Results: brightness-reversal test

Mean accuracy rates on the brightness-reversal test were slightly higher for normal than for reversed Greebles in all four group-test set conditions (overall means 0.88 for normal and 0.85 for inverted), but no main effects or interactions were significant in the ANOVA. Mean response times are plotted in Fig. 13. The main effect of group was not significant, but the main effects of test set and brightness were, $F(1, 20) = 19.3$ and

6.55, respectively. The interaction of interest, between group and brightness was also significant, $F(1, 20) = 5.57$. The source of this interaction can be seen in Fig. 13: experts were more impaired by brightness reversal than were novices. Post hoc LSD tests confirmed that experts were reliably slower (224 ms) at naming brightness-reversed Greebles than normal Greebles ($P < 0.005$) while the equivalent difference for novices (10 ms) was not reliable ($P = 0.88$). Although reversed Greebles from the heterogeneous set were especially difficult for experts to recognize, the three-way interaction of group, test set, and brightness was not significant, $F < 1$, so results from the two test sets are combined in Fig. 13.

6.4. Discussion

Results from the rotation and brightness-reversal tests supported our prediction that novices would more easily adapt to new stimulus presentation conditions than would experts. On the rotation test, the significant three-way interaction of group, block, and orientation indicates that novices, but not experts, became as fast at recognizing misoriented Greebles as they were at recognizing upright Greebles. That is, although experts were faster overall at recognizing Greebles, they remained relatively impaired on misoriented (compared to upright) Greebles even with practice; novices became as fast with misoriented as with upright Greebles. On the brightness-reversal test, experts but not novices were slower at recognizing brightness-reversed as compared to normal Greebles, as demonstrated by the significant two-way interaction of group and brightness.

Unlike in tests of the inversion effect with faces and other stimuli [6,36], experts did not show a large advantage in accuracy for upright over misoriented Greebles (or for normal over brightness-reversed Greebles). However, as previously suggested by Carey [37], the effect of mismatch in orientation between study and test is not the same as the difficulty in encoding inverted faces. Indeed, Ashworth and Tarr (in preparation) found that novice participants showed similar effects of rotation in recognizing both Greebles and faces that were learned upright.

Rather than comparing them to face inversion effects, the present findings are perhaps best considered in terms of canonicality effects in the object recognition literature [38]. The difficulty in naming any misoriented object depends, almost by definition, on participants having adopted a canonical orientation for an object class. The upright orientation was originally canonical for novices, but with a short amount of practice (block 1), the naming advantage for upright compared to misoriented Greebles disappeared in block 2. For experts, on the other hand, the canonicality of the upright orientation was much stronger (having been fostered over 9 h of training), leading performance on misoriented Greebles to be equally impaired in both blocks.

The implication of the present findings, that representations and/or processes used by experts are 'hyperspecific' with regard to novel stimulus presentation conditions, is intriguing considering that expertise does facilitate recognition of novel exemplars of a class. In other words, expertise generalizes to new members of a class, but does not always generalize to new stimulus conditions. This conclusion has important implications for models of expertise performance. Specifically, it may indicate that expertise is based on view- and image-specific representations of learned exemplars. Recognition using such representations would be particularly susceptible to image transformations such as orientation and brightness reversal.

A final point of discussion from Section 6 concerns the relative difficulty of misoriented and brightness-reversed Greebles for experts and novices. In recognizing brightness-reversed Greebles, experts were actually slower than novices, while for misoriented Greebles, experts were slightly faster than novices. One potential interpretation of these results is that experts tried to apply their expert processes to brightness-reversed Greebles and failed, while for misoriented Greebles, experts used their 'normal' object recognition processes, rather than relying on their expertise [11,39]. However, it could also be that the difference in patterns for the two tests is simply due to the fact that experts were not any better than novices at recognizing even the normal Greebles used in the brightness-reversal test (test sets C and D), while experts were better than novices at recognizing the upright Greebles in the rotation test (test set B). The conditions under which experts do and do not attempt to apply their expertise thus remains a question open to further study.

7. Modeling expertise

The expertise acquired by participants in this study and in Gauthier and Tarr's experiments with Greebles [4], as well as the expertise demonstrated by all human beings with faces, concerns individual-level recognition abilities. That is, Greeble experts are very good at distinguishing individual Greebles from each other (within the limits outlined in Sections 3 and 6), not just at distinguishing Greebles from other classes of objects. The latter task, class-level recognition of Greebles, is easily accomplished even by novices. Some researchers (e.g. Jolicoeur [40]) have proposed that separate object recognition systems are responsible for subordinateor individual-level and class-level recognition. In such a dual-system approach, the training in Section 2 would teach participants to use their subordinate-level recognition systems to recognize Greebles. An alternative

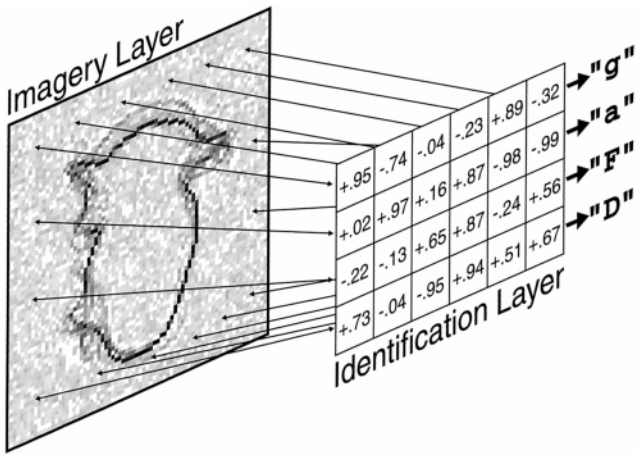

Fig. 14. Architecture of the WHOA model. Although all connections are not drawn, every unit in the imagery layer is connected to every unit in the identification layer, and vice-versa. In interpreting identification-layer patterns, values greater than 0 are treated as $+1$ and values less than 0 as -1, to produce an ASCII-like format.

possibility, which we endorse here, is that expert abilities in individual-level recognition can be accomplished through the same mechanisms that support novice, class-level recognition. If this hypothesis is correct, our experts' general-purpose object recognition system would simply have been 'tuned' to respond more precisely to Greebles than to other classes of objects.

As a demonstration of how a recognition system might show such tuning behavior, we simulated several aspects of the experiments in Sections 2–4 using a simple neural-network model originally presented by Williams [41]. The model, dubbed WHOA (for Widrow–Hoff Object Associator), is capable of learning both individual- and class-level information about objects via a single set of processes and representations. Although originally developed in connection with a completely different set of stimuli and psychophysical findings, we were able to apply the WHOA model, essentially without modification, to Greeble recognition.

7.1. Model architecture and simulation procedure

The basic architecture of WHOA is quite simple (Fig. 14), consisting of two layers of units that are fully connected both forwards and backwards. The imagery layer codes a 75×75-pixel image in 5625 U, while the identipcation layer codes arbitrary 24-unit patterns that can be associated with images. For the present simulations, the model was presented with black-and-white outline images of the Greebles (shown in Fig. 15) on the imagery layer.[6] Identification-layer patterns can be interpreted by using sets of six units to represent letters in an ASCII-like format (unit values are first 'discretized' to -1 or $+1$ for this interpretation). For the present simulations, the first letter was used to represent Greeble genders ('g' for GLIPs and 'P' for PLOKs), the second letter Greeble families ('a', 'g', 'j', 'p', or 'u' for each of the five families), and the third and fourth letters individual-level Greeble names (described below).

During training, connection weights are modified by a variant of the Widrow–Hoff learning rule (a.k.a. the delta or least-mean squares rule). Given an input pattern f and a target pattern t to be associated with the input, the change in weights ΔA is computed by the following formula:

$$\Delta A = \eta(t - \iota A f)f^{\mathrm{T}},$$

where η (eta) is a learning constant, ι (iota) is a 'generalization constant', and f^{T} is the transposition of f. When ι is 0, this formula reduces to a simple Hebbian learning rule, while an ι of 1 produces the standard Widrow–Hoff rule. For the present simulations, ι was set to 0.2; see [41] for a discussion of the model's behavior with various other values for ι.

The training procedure was intended to emulate the naming with feedback task, in which participants at-

[6] Interestingly, pilot simulations using greyscale images of the Greebles were not as successful as the simulations reported here using outline contours. The effectiveness of various image formats is a topic of current investigation with the WHOA model.

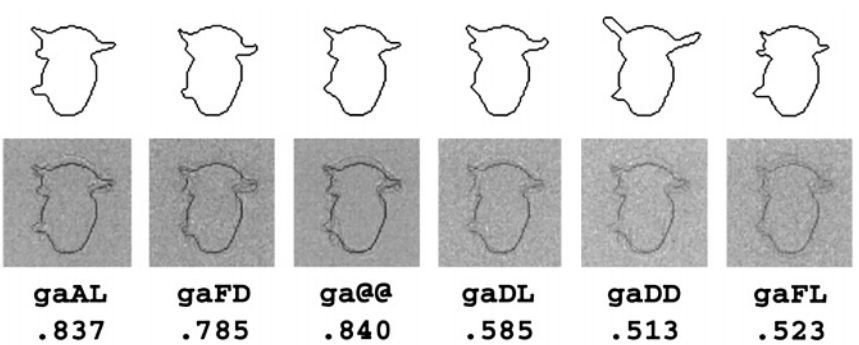

Fig. 15. WHOA's response to six test Greebles of the GLIP gender and 'a' family. The silhouettes presented to the model are shown in the top row, followed by the feedback-image, identification-layer label, and cosine responses to each Greeble.

tempted to name a Greeble, then saw the Greeble with its correct label if they were incorrect. Each simulated training trial involved two phases of weight-modification. In the first, the input image was fed through the forward connections to produce a guessed label, and both the forward and backward connections were altered based on this guess. For this phase of learning, η was set to 0.5, leading to incomplete association of the image with the guess. The second weight-modification phase depended on whether the guess was correct or not, where to be correct all four letters in the guessed label had to exactly match the correct label. If correct, weights were modified exactly as in the first phase, completing the learning of the image-guess association. If incorrect, weights were altered based on the correct label, with $\eta = 1.0$.

To test the model's performance, an image was presented on the imagery layer and fed through the forward connections to produce an identification pattern, then this identification pattern was fed through the backward connections to produce a 'feedback-image' pattern (connection weights were not altered during testing). There are two ways to evaluate performance. The simplest is to compare the letters in the interpreted identification pattern with the desired letters. This raw accuracy measure is somewhat crude (since information is lost when the identification-layer units are discretized for interpretation), but allows us to evaluate gender-, family-, and individual-level categorization performance independently (i.e. the model could get the gender letter correct but fail to produce the correct individual-name letters). A more complex measure involves computing the vector cosine of the observed feedback-image pattern and the original input pattern [42]. The cosine gives the best assessment of what WHOA 'knows' about a Greeble image, since it takes into account information stored in both the forward and backward connection weights.

7.2. Simulation training

The expertise training in Section 2 was simulated by teaching the model to associate the 30 training Greebles with appropriate labels. As in the human training, WHOA began by learning individual names (e.g. 'LR') for only five Greebles, and was initially taught to associate the other 25 Greebles with a NIL ('@@') label (the identification-layer pattern associated with each '@' character was $-1, -1, -1, -1, -1, -1$, and had to be learned by the model just like any other pattern). In three subsequent sets of trials, WHOA was taught new individual names for five Greebles that had been previously associated with the nil label (for example, a Greeble might have been associated with the label 'ga@@' in the first set of trials, but 'gaAK' in the second set). The final set of trials (in which 20 Greebles were named and ten unnamed) was then repeated twice. In each set of trials, the Greebles were presented in a different random order. Before every individual learning trial, WHOA was tested on the to-be-learned Greeble. Reminiscent of human participant's performance, WHOA was initially much better on unnamed than named trials for mean raw accuracy, but became essentially perfect (98% correct) on both types of trials by the end of training. However, the model's mean cosine for named Greebles was consistently higher than its mean cosine for unnamed Greebles.

7.3. Simulation tests

Following training, the model was tested on the 30 training Greebles and 30 additional Greebles that were not seen during training. The model always produced the correct gender, family, and individual letters to the trained Greebles. More interestingly, the model produced the correct gender and family letters for untrained Greebles with 97 and 95% accuracy, respectively. The high level of gender performance was especially surprising: WHOA learned the 'rule' for gender discrimination (whether boges, quiffs, and dunths point up or down) even though no two Greebles' parts were exactly the same. Furthermore, the model very rarely overgeneralized—in response to an unknown Greeble, it produced an individual label of a known Greeble only 1.3% of the time. Fig. 15 shows the model's response to six test Greebles of one gender and family. The three Greebles on the left were trained and the three on the right untrained. Note that the model produced very similar feedback-images to each of the untrained Greebles; this feedback-image could be considered the 'prototype' for all Greebles of this gender and family (see [41] for a detailed discussion of prototype effects in WHOA). Also note that the individual-level letters generated to the last Greeble, 'FL', are a combination of the first two trained Greebles' names, reflecting the fact that this Greeble's dunth was similar to to gaAL's, but its left boge and head shape was more similar to gaFD's.

WHOA was also trained to name the four novel Greeble test sets from Section 3 (Fig. 6). The model was given eight runs of trials with each test set, and training performance was evaluated as above. Performance by ten simulated experts was compared to performance by ten simulated novices, where the model was not given any training prior to learning the test sets (different simulation runs produced different results because of the random ordering of Greebles during learning trials). Raw accuracy was somewhat higher for the novice (73%, averaged over the last two training runs and all four test sets) than the expert model (67%), although the cosine measure was much higher for the expert than the novice model throughout training (first run: expert

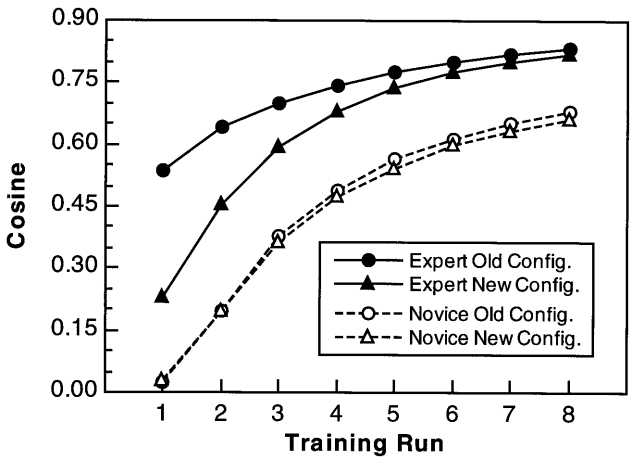

Fig. 16. Mean cosine measures for expert (solid lines) and novice (dashed lines) models learning Greebles in old (original; circles) and new (triangles) configurations.

0.42, novice 0.06; last run: expert 0.81, novice 0.63; averaged over all sets). Thus the expert model was not facilitated at learning names for new Greebles, but did learn considerably more information about Greeble images than the novice model. Raw accuracy was approximately equivalent on test sets A (77%, averaged over the last two training runs and over novices and experts) and B (69%), and was much worse on set D (38%), as was the case for human participants. However, the model found it very easy to learn the names of set C (96%), while human participants had more difficulty on this set than sets A and B. This result reflects the fact that the model was particularly good at distinguishing Greebles on the basis of body shape (the Greebles in set C all had different body shapes); apparently humans are not as good at utilizing this information. Interestingly, the cosine measure was essentially identical for the four sets in the expert model (0.83, 0.86, 0.80 and 0.80 in the last run of trials; for the novice model, cosines were 0.68, 0.73, 0.50, 0.71).

In a final simulation, expert and novice models were taught names for test set A Greebles in the new configurations used in Section 4 (only the most extreme configuration change was tested; Fig. 8 Version 3). Although this simulation was not directly analogous to the old/new configuration test, we expected the expert model to demonstrate quicker learning of Greebles in the old configuration than in the new configuration, if the model 'learned' anything about Greeble configurations during training. Fig. 16 shows mean cosines in each of the eight training runs for the old configurations (this is the data described in the last paragraph) and new configurations. Predictably, the novice model showed virtually identical training patterns for the two configurations; since this model knew nothing about Greebles before it began, it did not matter which configuration it learned. The expert model, however, responded much more strongly to Greebles in the old configuration, especially when it first began learning names for the set.

7.4. Discussion

In summary, the WHOA model was able to learn both individuating information and class-level information about Greebles, using a common set of units and connection weights. When trained on the 30 Greebles used in Section 2, the model showed performance patterns on named and unnamed items that were qualitatively similar to those of human participants. When tested on the 30 trained Greebles and on a new set of 30 untrained Greebles, WHOA always correctly recalled the gender, family, and individual names of trained Greebles, almost always correctly classified untrained Greebles by family and gender, and almost never confused trained with untrained inclividuals. Finally, when trained to learn names for new sets of Greebles, expert ancl novice models acquired the names at similar rates, but the expert model attained much higher cosines than the novice model, reflecting greater knowledge of the Greeble images themselves. This result is consistent with the fact that expertise effects were found with Test Sets C and D in the brightness-reversal test even though expert participants were not substantially faster than novices at acquiring names for these Greebles in Section 3.

The generalization ability of simple neural-network models has been documented in past reports by Knapp and Anderson [43] and McClelland and Rumelhart [44]. The novel aspect of WHOA is its ability to individuate trained Greebles at the same time that it learns to generalize to untrained Greebles. That is, the model acquires through training the ability to efficiently process novel instances of the Greeble class, but also retains the ability to distinguish among and provide labels for individual Greebles it has already learned. These abilities stem from a combination of two factors. First, the small amount of gradient-descent learning incorporated in the learning rule constantly pushes the network to respond most strongly to trained exemplars. Second, individuating information is explicitly stored (via the two-letter individual-level label) for each trained exemplar in its associated identification-layer pattern. The fact that the expert model attained higher cosines than the novice model on novel sets of test items indicates that WHOA's 'knowledge' of Greebles was not limited to the particular Greebles on which it was trained. Like our expert participants, the model also learned information that helped it to process new Greebles. Also like our expert participants, the model was not as good at learning names for Greebles with very subtle part differences (test set D), and was impaired in processing Greebles whose part-configurations

had been altered. While WHOA is only a demonstration model, not a full-fledged model of object recognition, its successes indicate that expert, individual-level recognition and novice, class-level recognition may be accomplished by the same cognitive system.

8. General discussion

We began this paper by asking whether the visual recognition mechanisms used by experts are the same as those used by novices. Although this question has been addressed in previous studies, they have typically used extant experts (e.g. face or dog experts), and therefore suffer from numerous limitations resulting from the unconstrained nature of expertise acquisition in everyday settings, such as the limited control over the training conditions and the confounding of perceptual and semantic learning (see [23] for one example of an experimental manipulation of expertise). In Section 2 of the present study, we showed how participants can be trained to become experts at recognizing individual exemplars of a novel class of objects (Greebles). Section 2 also demonstrated how performance on two tasks, Verification and Naming, changed as participants progressed from being Greeble novices to Greeble experts. In Sections 3–6, we compared the performance of the experts trained in Section 2 with the performance of novice participants on a variety of tasks designed to investigate: (1) the ways in which objects are represented and processed differently as a result of expertise, and (2) the extent to which expertise generalizes to new exemplars of an object class and to novel viewing conditions. The results of Sections 4 and 5 suggested that experts processed Greebles 'configurally' and 'holistically' [5], while in Sections 3 and 6, expertise mechanisms exhibited a surprising combination of generalizability to novel Greeble exemplars, but hyperspecificity for particular viewing conditions. In Section 7, we showed how a simple neural-network model (WHOA) could account for several aspects of both expert and novice Greeble recognition.

8.1. An expert is not an expert is not an expert

Although we (and other researchers) may have often discussed perceptual expertise as if it were a single, unified phenomenon, several aspects of the present results suggest that the participants trained in Section 2 acquired multiple skills over the course of becoming experts. First, consider the correlational data between the Verification and Naming tasks in Section 2. Intuitively, the two tasks, which involve matching a label to a picture and generating a label for a picture, respectively, seem quite similar to each other. In accordance with this intuition, participants' performance on the two tasks was, at the start of training, highly correlated across different Greebles, indicating that similar types of information were extracted from the images in order to perform the two tasks. However, over the course of the ten-session training procedure, this correlation went steadily down. Further investigation revealed a high correlation between Verification performance at the beginning and end of training, but a low correlation between Naming performance at the beginning and end of training (intriguingly, the WHOA model also showed a low item correlation on the cosine measure between the first and last set of training runs).

One possible interpretation of these findings is as follows. At the beginning of training, participants may have keyed in on individual Greeble parts to perform both the Verification and Naming tasks. In other words, participants might have identified a particular Greeble as 'Pimo' by its highly distinctive horse-like quiff, and 'Vali' by its dog-like boges. This strategy would continue to be effective on the Verification task throughout training, because every time a participant saw the label 'PIMO', she would know to look for the horse-like quiff in the subsequently-presented Greeble.[7] In the Naming task, however, participants are not given such a cue, so as more Greebles are learned, searching through the list of distinctive features would become more difficult. Instead of continuing with this strategy, participants might have learned to consider all parts together when performing the naming task. This hypothesis receives some support from Section 5, in which participants performed essentially the same naming task, but were asked to base their judgments on only the top or the bottom half of the test stimulus.

Experts were much more accurate on this task when the top and bottom halves came from the same Greeble than when the two halves came from different Greebles, even when the two halves were misaligned (Fig. 11). Thus being able to process Greebles 'holistically' (evaluating all the Greeble's parts together) aided experts in identifying an individual part even though the parts were not in the correct configuration.

Experts may have also learned to process Greebles 'configurally', although the effects supporting this conclusion in the present study only reached marginal significance levels (P values between 0.05 and 0.065). Configural processing can be seen as a more specific form of holistic processing, where the relations between parts are considered in addition to the parts themselves.[8] The first indication of configural processing

[7] Note that Gauthier and Tarr [4] found an old/new configuration advantage in experts following training which included only the Verification tast, indicating that even if participants do use distinctive features in Verification, the task must also encourage acquisition of configural processing abilities.

[8] Note that what Carey and Diamond [5] call 'holistic processing' implies both mechanisms discussed here. For further discussion of the various definitions of configural and holistic processing, see [5,60].

comes from Section 4, where experts but not novices were more accurate in identifying a Greeble's quiff when it was presented in the context of the learned configuration than in a new configuration (Fig. 9). Gauthier and Tarr [4], using a paradigm that matched more closely that previously used with faces [18], found the same old/new configuration advantage with all three Greeble parts (quiffs, dunths, and boges), and again obtained reliable effects with Greeble experts but not with novices. The second result consistent with configural processing involves trials from the composite task of Section 5 in which the top and bottom halves of displays came from the same Greeble. Experts were faster and more accurate when the two halves of the original Greeble were aligned than when the two halves were misaligned; misalignment obviously alters the configuration of the Greeble parts. Finally, configural processing may also be the mechanism by which participants learned to perform the Verification task as quickly with individual labels as with gender labels (Figs. 4 and 5), an effect that Gauthier and Tarr [4] gave as their criterion for expertise status.

8.2. Relevance of findings to the face-specificity issue

The issue of whether faces are 'special' is both complex and independent from the issue of whether expertise effects in visual recognition can be found for non-face object classes. Admittedly, the results presented here, by themselves, do not provide compelling evidence against face-specific processing. While our Greeble experts display some of the putatively face-specific behavioral effects (configuration effects on quiffs, composite effect, contrast inversion), they fail to display others (orientation inversion effect, configuration effect on parts other than quiffs). Clearly, the 9 h or so of training experienced here did not lead our subjects to process Greebles in exactly the same manner as they process faces, for which young adults have had approximately 20 years of training. On the other hand there is some evidence that far less experience may be needed to show configural effects with faces. Tanaka et al. [45] have recently demonstrated that children as young as 5 years of age are 'face experts' to the extent that they process upright faces configurally. Of course, the five years of experience that young children have with faces is still substantially more than the training that our experts received with Greebles. Thus, the ability to use configural coding may develop only slowly, so that some behavioral effects found with non-face objects may not reach magnitudes comparable to those found with faces without an equivalent experience with the alternate class. Nevertheless, evidence that any behavioral effect can be obtained with at least one class of non-face stimuli in at least one context should be sufficient to discredit the effect as evidence for face-specific processing. Results from the present study suggest that the contrast inversion effect and composite effect should no longer be considered face-specific. Orientation inversion effects [6], old/new configuration effects [4], and whole/part advantage effects TaGa97 have been found for alternate stimulus classes elsewhere.

Additional evidence for face-specific processing comes from neuroimaging [46–51] and neuropsychological studies [39,52–54], which our present results do not address. Even if faces and other objects are considered to be processed by similar mechanisms, whether any brain area is or is not face-specific remains an empirical question [55]. To date results on this issue have been mixed, with some studies indicating that an area in the inferior temporal cortex is preferentially activated by faces [48,56] and other studies suggesting that this same area can be engaged by within-class recognition of common objects, such as identifying a car as a Ferrari or a Honda [57]. Finally, some of us have recently collected evidence using fMRI which is particularly relevant to the present study—Gauthier et al. [57,58] have found that the face area of individuals undergoing expertise training with Greebles becomes increasingly activated when discriminating amongst upright (as opposed to inverted) Greebles. Thus, there is some evidence that expertise significantly modifies the neural substrates engaged during object recognition. The simulation data presented in Fig. 16 indicates that something analogous happens in the neural-network WHOA model when it is 'trained': a subset of the model's units are altered in such a way that Greebles presented in normal configurations are processed differently than Greebles presented in altered configurations. With this evidence in mind, it is crucial to understand the perceptual mechanisms underlying this change. The results presented here provide a start in this direction, helping to elucidate the cognitive structures and processes that support expert visual recognition.

Acknowledgements

The contribution of the first two authors was equal. This research was supported by Air Force Office of Scientific Research Grant F49620-91-J0169 and National Science Foundation Grant SBR-9615819 to MJT, and by NIH Grant R15 DH30433 to JT. Thanks to Seth Koplowitz for help in collecting data for Section 2.

References

[1] Farah MJ. Is an object an object an object? Cognitive and neuropsychological investigations of domain-specificity in visual object recognition. Curr Directions in Psychol Sci 1992;1(5):164–9.

[2] Tanaka JW, Gauthier I. Expertise in object and face recognition. In: Goldstone RL, Schyns PG, Medin DL, editors. Psychology of Learning and Motivation, vol. 36. San Diego, CA: Academic Press, 1997:83–125.
[3] Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. Basic objects in natural categories. Cognit Psychol 1976;8:382–439.
[4] Gauthier I, Tarr MJ. Becoming a 'Greeble' expert: Exploring the face recognition mechanism. Vis Res 1997;37(12):1673–82.
[5] Carey S, Diamond R. Are faces perceived as configurations more by adults than by children? Visual Cognition 1994;1(2/3):253–74.
[6] Diamond R, Carey S. Why faces are and are not special: An effect of expertise. J Exp Psychol: General 1986;115(2):107–17.
[7] Tanaka JW, Taylor M. Object categories and expertise: Is the basic level in the eye of the beholder? Cognit Psychol 1991;23:457–82.
[8] Edelman S. Representation, similarity, and the chorus of prototypes. Minds and Machines 1995;5(1):45–68.
[9] Myles-Worsley M, Johnston WA, Simons MA. The influence of expertise on X-ray image processing. J Exp Psychol: Learning Memory and Cognition 1988;14(3):553–7.
[10] Rhodes G, Tan S, Brake S, Taylor K. Expertise and configural coding in face recognition. Br J Psychol 1989;80:313–31.
[11] Farah MJ, Wilson KD, Drain HM, Tanaka JW. The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. Vis Res 1995;35:2089–93.
[12] Scapinello KF, Yarmey AD. The role of familiarity and orientation in immediate and delayed recognition of pictorial stimuli. Psychon Sci 1970;21:329–31.
[13] Yin RK. Looking at upside-down faces. J Exp Psychol 1969;81(1):141–5.
[14] Rhodes G, Brake S, Atkinson AP. What's lost in inverted faces? Cognition 1993;47(1):25–57.
[15] Rock I. The perception of disoriented figures. Sci Am 1974;230:78–86.
[16] Sergent J. An investigation into component and configural processes underlying face perception. Br J Psychol 1992;75(2):221–42.
[17] Tanaka JW, Farah MJ. Parts and wholes in face recognition. Q J Exp Psychol 1993;46A:225–45.
[18] Tanaka JW, Sengco JA. Features and their configuration in face recognition. Mem Cognit 1997;25:583–92.
[19] Tanaka JW, Giles M, Szechter L, Lantz JA, Stone A, Franks L, Vastine K. Measuring parts and wholes recognition of cell, car, and dog experts: A test of the expertise hypothesis. Unpublished Manuscript, Oberlin College, 1997.
[20] Schyns PG, Murphy GL. The ontongeny of part representation in object concepts. In: Medin D, editor. The psychology of learning and motivation, vol. 31. San Diego, CA: Academic Press, 1994:305–54.
[21] Schyns PG, Goldstone RL, Thibaut J-P. The development of features in object concepts. Behavioral and Brain Sciences 1997:in press.
[22] Ashby FG, Maddox WT. A response time theory of separability and integrality in speeded classification. J Math Psychol 1994;38:423–66.
[23] Ashby FG, Maddox WT. Complex decision rules in catergorization: Contrasting novice and experienced performance. J Exp Psychol: Human Perception and Performance 1992;18:50–71.
[24] Young AW, Hellawell D, Hay D. Configural information in face perception. Perception 1987;10:747–59.
[25] Hole GJ. Configurational factors in the perception of unfamiliar faces. Perception 1994;23(1):65–74.
[26] Kohler W. Dynamics in Psychology. New York: Liveright, 1940.
[27] Tarr MJ, Pinker S. Mental rotation and orientation dependence in shape recognition. Cognit Psychol 1989;21:233–82.
[28] Tarr MJ, Gauthier I. Do viewpoint-dependent mechanisms generalize across members of a class? Cognition 1997: submitted.
[29] Gauthier I, Tarr MJ. Orientation priming of novel shapes in the context of viewpoint-dependent recognition. Perception 1997;26:51–73.
[30] Moses Y, Ullman S, Edelman S. Generalization to novel images in upright and inverted faces. Perception 1996;25:443–62.
[31] Carey S. Perceptual classification and expertise. In: Gelman R, Kit-Fong T, editors. Perceptual and Cognitive Development. San Diego, CA: Academic Press, 1996:49–69.
[32] Galper RE. Recognition of faces in the photographic negative. Psychon Sci 1970;19:207–8.
[33] Hayes T, Morrone MC, Burr DC. Recognition of positive and negative bandpass-filtered images. Perception 1986;15:595–602.
[34] Johnston A, Hill H, Carman N. Recognising faces: Effects of lighting direction, inversion, and brightness reversal. Perception 1992;21:365–75.
[35] Phillips R. Why are face hard to recognize in photographic negative? Percept Psychophys 1972;12:425–6.
[36] Bruyer R, Crispeels G. Expertise in person recognition. Bull Psychon Soc 1992;30(6):501–4.
[37] Carey S. Becoming a face expert. Philos Trans R Soc London, B 1992;335:95–103.
[38] Palmer S, Rosch E, Chase P. Canonical perspective and the perception of objects. In: Long J, Baddeley A, editors. Attention and Performance ix. Hillsdale, NJ: Lawrence Erlbaum, 1981.
[39] Farah MJ, Levinson KL, Klein K. Face perception and within-category discrimination in prosopagnosia. Neuropsychologia 1995;33:661–74.
[40] Jolicoeur P. Identification of disoriented objects: A duel-systems theory. Mind and Language 1990;5(4):387–410.
[41] Williams P. Prototypes, exemplars, and object recognition. Boston: University of Massachusetts, 1997, submitted.
[42] Jordan MI. An introduction to linear algebra in parallel distributed processing. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. Cambridge, MA: MIT Press, 1986:170–215.
[43] Knapp AG, Anderson JA. Theory of categorization based on distributed memory storage. J Exp Psychol: Learning, Memory and Cognition 1984;10:616–37.
[44] McClelland JL, Rumelhart DE. A distributed model of human learning and memory. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 2. Cambridge: MA: MIT Press, 1986:170–215.
[45] Tanaka JW, Kay JB, Grinnell E, Stansfield B, Szechter L. Face recognition in young children: When the whole is greater than the sum of its parts. Visual Cognition 1997:in press.
[46] Haxby JV, Horwitz B, Ungerleider LB, Maisog JM, Pietrini P, Grady CL. The functional organization of human extrastriate cortex: A pet-rcbf study of selective attention to faces and locations. J Neurosci 1994;14:6336–53.
[47] Kanwisher N, Chun MM, McDermott J, Ledden PJ. Functional imaging of human visual recognition. Cognit Brain Res 1996;5:55–67.
[48] Kanwisher N, McDermott J, Chun MM. The fursiform face area: A module in human extrastriate cortex specialized for face perception. J Neurosci 1997:in press.
[49] Puce A, Allison T, Gore JC, McCarthy G. Face-sensitive regions in human extrastriate cortex studied by functional mri. Neurophysiology 1995;74(3):1192–9.
[50] Sergent J, Ohta S, Signoret J-L. Functional neuroanatomy of face and object processing. Brain 1992;115':15–36.
[51] Sergent J, Signoret J-L. Functional and anatomical decomposition of face processing: Evidence from prosopagnosia and pet study of normal subjects. Philos Trans R Soc Lond B 1992;335:55–62.

[52] Sergent J, Signoret J-L. Varieties of functional deficits in prosopagnosia. Cerbral Cortex 1992;2:375–88.
[53] McNeil JE, Warrington EK. Prosopagnosia: A face-specific disorder. Q J Exp Psychol 1993;46A(1):1–10.
[54] Moscovitch M, Winocur G, Behrmann M. What is special in face recognotion? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. J Cognit Neurosci 1997:in press.
[55] Hay DC, Young AW. The human face. In: Ellis AW, editor. Normality and patholgy in cognitive function. London: Academic Press, 1982.
[56] McCarthy G, Puce A, Gore JC, Allison T. Face-specific processing in the human fusiform gyrus. J Cognit Neurosci 1997:in press.
[57] Gauthier I, Anderson AW, Tarr MJ, Skudlarski P, Gore JC. Levels of categorization in visual object studied with functional MRI. Curr Biol 1997;7:645–51.
[58] Gauthier I, Tarr MJ, Anderson A, Skudlarski P, Gore J. Expertise training with novel objects can recruit the fusiform face area. In: Society for Neuroscience Ahstract. New Orleans, LA, 1997.
[59] Loftus GR, Masson MEJ. Using confidence intervals in within-subject designs. Psychon Bull Rev 1994;1:476–90.
[60] Farah MJ, Tanaka JW, Drain HM. What causes the face inversion effect? J Exp Psychol: Human Perception and Performance 1991;21(3):628–34.